前 言

数值分析的主要目标是仅用最简单的算术运算来获得复杂问题的近似解.换句话说,它的任务是用一系列容易的步骤来解困难的问题.显然,这意味着寻求可在计算机上使用的算法步骤来为我们解题.问题来自数学的方方面面,特别是代数与分析,彼此间的界限有时是难以分清的.数值分析家从这些领域中借用了不少背景理论,为明晰起见,某些内容还必须包含在入门课程中.与此同时,我们的主题也打破界限反馈不少素材.数值方法已对代数和分析理论作出重要贡献.

在这第二版中加进了许多新的课题,包括向后分析、样条、自适应积分、快速 Fourier 变换、有限元、刚性微分方程以及 QR 方法.线性方程组那一章已经完全改写.一些老课题被缩减或取消.然而由于历史的原因古典数值分析中有代表性的内容仍部分地保留.某些割舍使作者甚感惋惜,尤其痛心的是删除了微分方程解的存在性的构造性证明.总体来说,新版内容更符合客观要求,但同样可以说是课程本身的需要.

它的陈述方式与目的保留不变.其材料适用于第一年学位课程.在作恰当的筛选后也可方便地作为一个学期的入门课程.题目的叙述格式既可方便地作为其他课本的补充,也可作为独立的研究.每一章开头依然是本章内容的综述,并且可以看作是本章的内容索引.其内容不打算靠其自身来说明,而是在解题中提供它们的细节.

下面重述我在第一版前言中的最后一段话:毫无疑问,虽然已竭尽全力,而错误仍在所难免.也许由于犯错误的机会是如此之多,以致于数值分析工作者属于世界上最能自我意识到出错的一类人.对于来自读者的发现错误的声音我是由衷地欢迎.(对第一版中这一吁请响应者寥寥.)除了对所有难以理解的"真理"作共同的探讨所感到的快慰之外,别无酬谢.

F.施依德

目 录

第一章　数值分析是什么

算法

数值分析的目的是仅用简单的算术运算解复杂的数值问题，并开发和评估由给出的数据计算数值结果的方法. 这些计算方法被称为算法.

我们将致力于对算法的研究. 迄今为止，仍未找到某些问题的满意算法；另一方面，当存在着几种不同算法时，我们必须从中作出选择. 选择这个而非那个算法时存在种种理由，但有两个明显的标准：速度与精度. 速度快显然是一种优势，虽然就适当规模的问题而言，这一优势因计算机的能力而被削弱殆尽. 但对于大规模的问题，速度仍是重要的因素，速度慢的算法由于不实用会被淘汰. 然而，在其他条件均相同时，速度较快的方法定会被首肯.

例 1.1　求出 2 的平方根，直至具有 4 位十进制小数.

仅用 4 种基本的算术运算，存在着不只一个算法. 无疑，算法

$$x_1 = 1, \quad x_{n+1} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right)$$

是令人满意的. 据此，由少数几步心算就能很快得到

$$x_2 = \frac{3}{2}, \quad x_3 = \frac{17}{12}, \quad x_4 = \frac{1}{2}\left(\frac{17}{12} + \frac{24}{17}\right),$$

或四舍五入到 4 位十进制小数，

$$x_2 = 1.500, \quad x_3 = 1.4167, \quad x_4 = 1.4142.$$

最后的 x_4 所有的 4 位小数均是正确的. 这个数值算法具有悠久的历史，并且，它将在第 25 章中作为方程求根的特殊情况而再次遇到.

误差

数值计算的乐观主义者会问计算结果有多精确；而数值计算的悲观主义者会问结果中已引入了多少误差. 显然，这两个问题是相同的. 给出的数据，很少是精确的，因为它通常源于测量过程. 从而，输入的信息中或许存在着误差. 并且，算法本身通常也会带来误差，或许是不可避免的舍入误差. 这样，输出的信息中将包含出自这两种来源的误差.

例 1.2　假设数 0.1492 准确到给出的 4 位十进制小数，换言之，它是处在 0.14915 与 0.14925 之间的一个真值的近似，那么，其误差至多为第 5 位小数的 5 个单位或第 4 位小数的半个单位. 在此情况不，这个近似值被称为具有 4 位有效数字. 类似地，倘若 14.92 的误差不超过 0.005，则它具有 2 位准确的小数和 4 位有效数字.

例 1.3　当数 0.10664 缩写成 0.1066 时，被称为四舍五入成 4 位小数. 而 0.10666 将被四舍五入成 0.1067. 若给出的数字是准确的，则在这两种情况下，由舍入造成的误差不大于 0.00005. 前一个例子是向下"舍"，后一个例子是向上"入". 像 0.10665 这种边界状况，通常舍入到最接近的偶数数字，这里舍到 0.1066. 这避开了长期以来不知向下舍还是向上入的尴尬.

例 1.4　1.492 乘以 1.066，积是 1.590472. 计算机按固定的"字长"工作，所有的数均剪裁成这个长度. 假设有一台虚构的 4 位数字的机器，那么，上述的积将被四舍五入到 1.590. 这种舍入误差是算法误差，由现代计算中不可避免的大量计算所造成.

支撑理论

虽说我们数值分析的着眼点将面向应用，但我们会自然而然地涉及支撑理论（supporting

theory). 这种理论用于发现算法及建立算法的有效性. 通常指导我们的这个理论具有本质的趣味;它是有魅力的数学. 于是,我们就有了两个最好的领域. 但切勿忘记,我们的兴趣更多的应是实用性的而不是纯美学的.

例 1.5　计算三角函数、指数函数以及其他非初等函数的值,显然要依赖支撑理论. 对于小的 x, 为了获得它的余弦值,经典的级数

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$$

仍是一个好选择. 当 $x = 0.5$,它变成

$$\cos 0.5 = 1 - 0.125 + 0.0026041 - 0.0000217 + \cdots = 0.877582.$$

这是足够精确的. 在此情况下,更深层的支撑理论可确定误差界:该理论指出,对于像这样的级数,误差不大于第一个略去的项(见题 1.9). 这里,第一个略去的项是$\frac{x^8}{8!}$,对于 $x = 0.5$,它恰好达到小于 0.0000001.

数的表示法

由于我们的最终目标是数值的,故关于数的表示法,仅用一两句话将是不适当的. 由于我们最熟悉十进制的数,故数值输入通常将使用这种形式. 然而,几乎众所周知,通常计算机发现二进制表示法更方便. 它的 0 和 1 对应于电路的关与开或电压的高与低状态. 对于正整数,二进制的形式是

$$d_n 2^n + d_{n-1} 2^{n-1} + \cdots + d_1 2^1 + d_0 2^0,$$

而对小于 1 的正数,它是

$$d_{-1} 2^{-1} + d_{-2} 2^{-2} + d_{-3} 2^{-3} + \cdots,$$

其中,所有的二进制数字 d_i 取 0 或 1. 这种表示是惟一的.

浮点表示法(floating-point representation)格外方便. 在这种形式中,数字用三个部分来描述:一是符号,一是尾数(mantissa),一是阶(exponent)(自身也带有一个符号). 作为最初的例子,回到十进制,数 0.1492 能表示为

$$+0.1492\ 10^0$$

它的符号是 +,尾数是 0.1492,而阶是 0. 在其他的可能性中,以 $+1.492\ 10^{-1}$ 代替它也是可以的,但一般的习惯要求第一位(非零)数字恰好出现在小数点后,然后,由阶来处理数量级. 这种表示法被称为规格化(normalized). 于是,1492 将被表示为 $0.1492\ 10^4$.

例 1.6　将十进制的 13.75 转换为二进制的浮点形式.

有更正式的转换方法. 但即使没有它们,由于小数点的左边是 8 + 4 + 1,而右边是 $\frac{1}{2} + \frac{1}{4}$,易见 13.75 等价的二进制数是 1101.11. 现将它写成

$$+0.110111(+100),$$

其中,圆括号中的 100 起着阶 4 的作用,最终转换成

$$01101110100.$$

倘若懂得某种约定,对于电的效果来说,这数中仅有 0 和 1 是令人感兴趣的. 首位 0 被解释为正号(1 将意味着负号). 然后,6 个二进制数字或比特(bit)作为尾数,并假定在该尾数前有一个二进制的小数点. 紧接着的 0 是另一个正号,它是阶的符号,然后,这个阶结束该表示法. 最终的形式看上去很不像 13.75,但它能被理解. 实际中,尾数与阶将包含更多的位数,并且符号与阶的形式也将产生变化,而浮点表示法是现代计算中的一个基本工具.

向量与矩阵的范数

一个向量的 Euclid 长度,对于分量为 v_i 的向量 V 来说,即

$$(v_1^2 + v_2^2 + \cdots + v_n^2)^{1/2},$$

它也被称为 V 的一个范数(norm),以符号 $\| V \|$ 记之.这个范数 的 3 个基本性质是

1. $\| V \| \geqslant 0$, $\| V \| = 0$ 当且仅当 $V = 0$;
2. $\| cV \| = |c| \, \| V \|$,对任意常数 c;
3. $\| V + W \| \leqslant \| V \| + \| W \|$.

最后一个式子即通常所说的**三角不等式**.

若干其他的实函数也具有这些性质,并且也被称为范数.令人特别感兴趣的是 L_p 范数

$$\| V \|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}}, \quad p \geqslant 1.$$

当 $p=1$ 时,它是 L_1 范数,是各分量的绝对值之和.当 $p=2$ 时,它就是熟知的向量长度或Euclid范数.当 p 趋向于无穷大时,取出占优势的 v_i,于是我们得到最大范数(maximum norm)

$$\| V \|_\infty = \max_i |v_i|.$$

在不止一个场合,尤其是在研究算法的误差特性时,我们将能找到这些范数的用途.

例 1.7 利用 L_1 范数,向量(1,0),$\left(\frac{1}{2},\frac{1}{2}\right)$,(0,1)中的每一个均有范数 1.这种单位向量的平面图给在图 1.1(a)中.原点是它们的起点,而它们的终点构成一个正方形.图 1.1(b)表示更为常见的Euclid 范数下的单位向量.利用 L_∞范数,向量(1,0),(1,1),(0,1)中的每一个均有范数 1,它们的平面图如图 1.1(c)所示,它们的终点也构成一个正方形.

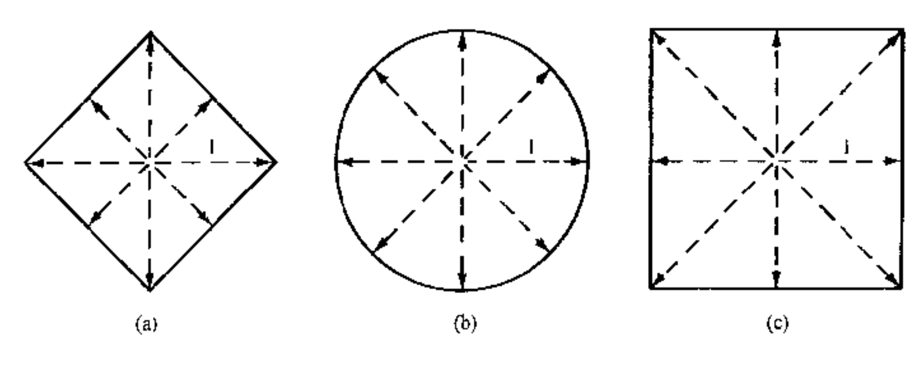


图 1.1

转到矩阵,我们定义

$$\| A \| = \max \| AV \|,$$

这个最大值是取遍所有的单位向量 V 而得到的.这里,"单位"的含义取决于所用的向量范数的类型.这种矩阵范数有平行于上文对向量列出的那些性质:

1. $\| A \| \geqslant 0$, $\| A \| = 0$ 当且仅当 $A = 0$;
2. $\| cA \| = |c| \, \| A \|$,对于任意常数 c;
3. $\| A + B \| \leqslant \| A \| + \| B \|$.

此外,对于矩阵 A、B 与向量 V,还有如下有用的性质:

4. $\| AV \| \leqslant \| A \| \, \| V \|$;
5. $\| AB \| \leqslant \| A \| \, \| B \|$.

L_1 范数和 L_∞范数具有容易计算的优点,前者是具有最大绝对值的一列之和,

$$\| A \|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|,$$

而后者是具有最大绝对值的一行之和,

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|.$$

它们的许多特性将在题解中被证明.

例 1.8 找出矩阵

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

的 L_1, L_2 与 L_∞ 范数.

能立即找出最大的列和及行和,于是我们从

$$L_1 = L_\infty = 2$$

迅速着手.不幸的是,没有相应的支撑理论对 L_2 提供帮助以及无困难地使这一外表十分简单的矩阵获得 L_2 范数的值.根据定义,A 的 L_2 范数是向量

$$\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x + y \\ x \end{pmatrix}$$

的最大 L_2 范数,其中 $x^2 + y^2 = 1$,即(x, y)在图 1.1(b)的单位圆周上.这个范数的平方是

$$(x+y)^2 + x^2 = 1 + 2xy + x^2 = 1 + 2x\sqrt{1-x^2} + x^2.$$

根据基本的微积分知识能取到它的最大值.在这里,由于对(x, y)与$(-x, -y)$,范数取值相同,故不妨设 y 是正的.最后得到:当 $x^2 = \frac{1}{2} + \frac{\sqrt{5}}{10}$时,有最大值

$$\|A\|_2^2 = \frac{3+\sqrt{5}}{2}.$$

题　　解

1.1 计算多项式

$$p(x) = 2x^3 - 3x^2 + 5x - 4$$

在自变量 $x = 3$ 时的值.

解 ☞ 依照自然过程,我们求出 $x^2 = 9, x^3 = 27$,然后将它们合在一起,

$$p(3) = 54 - 27 + 15 - 4 = 38,$$

共计执行了 5 次乘法,1 次加法,2 次减法.

现将该多项式重新整理为

$$p(x) = [(2x - 3)x + 5]x - 4,$$

并且再试算一次.从 $x = 3$ 开始,我们相继得到 6, 3, 9, 14, 42 与 38,这回只用了 3 次而不是 5 次乘法.减少量虽不惹人注意,但它却具有启发意义.对于一个一般的 n 次多项式,第一个算法需要 $2n-1$ 次乘法,而第二个算法仅需要 n 次算法.在一个较大的运算中,包含着许多多项式的求值运算,减少时间与算法(舍入)误差会意义重大.

1.2 定义一个近似值的误差.

解 ☞ 传统的定义是

$$\text{真值} = \text{近似值} + \text{误差}.$$

因此,例如,

$$\sqrt{2} = 1.414214 + \text{误差},$$
$$\pi = 3.1415926536 + \text{误差}.$$

1.3 相对误差是什么?

解 ☞ 相对误差(relative error)是相对于真值所度量的误差:

$$\text{相对误差} = \frac{\text{误差}}{\text{真值}}.$$

通常的情况下，真值是未知的或者是不适用的，稍加放宽，用近似值代替它并将该结果仍称为相对误差．于是，对于$\sqrt{2}$来说，熟知的近似值 1.414 有大约为

$$\frac{0.0002}{1.414}\approx 0.00014$$

的相对误差，而更粗糙的近似值 1.41 有一个接近 0.003 的相对误差．

1.4 设数 $x_1, x_2, \cdots, x_n$ 分别是 $X_1, X_2, \cdots, X_n$ 的近似值，而各自最大的可能误差是 E，证明 x_i 之和的最大的可能误差是 nE．

证 ☞ 因为

$$x_i - E \leqslant X_i \leqslant x_i + E,$$

由加法得到

$$\sum x_i - nE \leqslant \sum X_i \leqslant \sum x_i + nE,$$

因此

$$-nE \leqslant \sum X_i - \sum x_i \leqslant nE.$$

这就是所要证的．

1.5 计算$\sqrt{1}+\sqrt{2}+\cdots+\sqrt{100}$的和，其中，所有的平方根计算到小数点后两位．按照上题，最大的可能误差是多少？

解 ☞ 由适当选择的很少几行程序，或更老式地，求助于表格，能得到这个问题中的各个根，然后求和，其结果是 671.38．由于每个根有一个最大的可能误差 $E=0.005$，所以，和的最大的可能误差 $nE=100(0.005)=0.5$．这意味着所得的和也许连一位准确的小数也没有．

1.6 一个计算结果的概率误差(probable error)的意思是什么？

解 ☞ 这是一种使实际误差超过估计值有$\frac{1}{2}$概率的误差估计．换言之，实际误差可能大于或小于估计．由于它取决于误差的分布，所以不是一个容易研究的对象，而通常它被$\sqrt{n}E$ 粗略代替．其中，E 是最大的可能误差．

1.7 题 1.5 的结果中实际误差是多少？将它与最大误差及概率误差相比较，结果会怎样？

解 ☞ 当平方根求到有 5 位小数时，新的计算方法以和 671.46288 为结果．此时，最大误差是 100(0.000005)，即 0.0005．从而我们有准确到 3 位小数的和 671.463．于是，题 1.5 的结果中实际误差大约是 0.08，与最大(的可能)误差 0.5 及概率误差 0.05 相比，我们的估计一个太悲观而另一个又有些乐观．

1.8 1000 个而不仅仅是 100 个平方根相加，若想得到 3 位小数精确度，那么，参与计算的各个平方根应精确到何种程度？

解 ☞ 为了确保精度，最好是假设最坏的结果即最大的可能误差能达到．题 1.4 中的公式 nE 变成 $1000E$，这指出了在这种长度的一个求和中，可能会失去 3 位十进制小数．由于希望在输出中有 3 位准确小数．所以在输入中具有 6 位准确小数也许是明智的．其要点是，在一个长长的计算中，存在着非常小的误差汇聚成大值的机会．

1.9 计算级数

$$1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\cdots,$$

准确到 3 位数字．

解 ☞ 这个级数描述了一个常用的分析定理：因为它的项在符号上交替出现正、负，并且不断变小，其部分和来回穿越其极限即该级数的值，这意味着在任一点处的误差将小于第一个舍去的项．为了获得指定的精度，我们因此需要$\frac{1}{n}\leqslant 0.0005$ 或 $n\geqslant 2000$．必须将这 2000 项相加．对 8 位十进制小数运算，2000 个舍入误差可能会积累到 $nE=2000(0.000000005)=0.00001$．这看上去是微不足道的，从而我们允许计算进行下去，将结果舍到 3 位，有 0.693．

注意，在这个问题中没有输入误差，只有算法误差．首先，我们仅用一个部分和代替这个级数，然

后，在试图求这个和的值时，我们制造了大量的舍入误差．前者被称为截断误差（truncation error），并且，在这个问题中，它看上去是两个误差来源中较大的一个．概言之，

$$\text{实际误差} = \text{截断误差} + \text{舍入误差} = 0.0005 + 0.00001.$$

实际上，这个级数的值是 2 的自然对数，而取 3 位小数，它就是我们的 0.693．

1.10 证明：若级数

$$a_1 - a_2 + a_3 - a_4 + \cdots$$

是收敛的，而且，所有的 a_i 均为正数，则

$$\frac{1}{2}a_1 + \frac{1}{2}(a_1 - a_2) - \frac{1}{2}(a_2 - a_3) + \frac{1}{2}(a_3 - a_4) + \cdots$$

也是收敛的，并且表示相同的数．

证 ☞ 用 A_n 和 B_n 表示上述两个级数的 n 项部分和．易见，$A_n - B_n = \pm\frac{1}{2}a_n$．由于前一个级数是收敛的，$\lim a_n = 0$，从而得到结果．

1.11 将前题中的定理用于求题 1.9 中的级数值，仍准确到 3 位小数．

解 ☞ 稍用一点代数知识就可得出 $B_1 = \frac{1}{2}$，且对于 $n > 1$，

$$B_n = \frac{1}{2} + \sum_{k=2}^{n}(-1)^k \frac{1}{2k(k-1)}.$$

这仍是一个具有单调下降项的交错级数，从而题 1.9 中的定理仍有效．为了 3 位数字精确，我们需要

$$\frac{1}{2n(n+1)} \leqslant 0.0005,$$

或 $n \geqslant 32$．这远远少于题 1.9 中所需的项数，因而在一个 8 位十进制的机器中，舍入误差简直就不成为一个问题了．新算法较之另一个快多了，并且，它用较少的工作量就获得了相同的 0.693．

1.12 给出足够准确的数 0.1492 和 0.1498，即，其误差不大于第 5 位小数的 5 个单位．根据所考虑的商 $\frac{1}{0.1492 - 0.1498}$ 来阐明相对误差的形成．

解 ☞ 对于这些给出的数来说，相对误差约为 $\frac{5}{15000}$，接近 0.03 个百分点．对于它们的和与差而言，产生第 4 位小数中一个单位的最大误差是可能的．在和的情况下，我们仍导出一个大约为 0.03 个百分点的相对误差，但对于差 0.0006 而言，我们得出一个达到 $\frac{1}{6}$ 的误差，这是 17 个百分点．回到所求的商，它也许恰好达到最糟的情况．如所给的，取最接近的整数，算出的商将是 1667．可设想它本该是 $\frac{1}{0.14985 - 0.14915}$ 但却被取代了，而这使我们得到 1429．另一个极端是 $\frac{1}{0.14975 - 0.14925} = 2000$．这个非常简单的例子清楚地说明，在某些持续计算的内部过程中，一个大的相对误差完全可能引出大的绝对误差．

1.13 数值问题的条件指的是什么？

解 ☞ 如果输入信息中的小变化只引起输出的小变化，则称该问题是良态的（well-conditioned）；反之，它是病态的（ill-conditioned）．例如，系统

$$x + y = 1,$$
$$1.1x + y = 2$$

呈现出一个明显的难点：它表示几乎平行的两直线的交叉，而解为 $x = 10, y = -9$．

今将 1.1 改为 1.05 然后再求解．此时 $x = 20$ 而 $y = -19$．一个系数上的 5 个百分点的改变导致了解的 100 个百分点的改变．

1.14 什么是**稳定的**算法？

解 ☞ 在持续的计算中，将可能产生许多舍入误差．这些舍入误差中的每一个都扮演着剩下的计算中输入误差的角色，并且都对最终的输出结果产生影响．若所有这些误差对算法的累积影响是有限的，算法能获得有用的结果，则称之为**稳定算法**（stable algorithms）．不幸的是，有这种时候：误差的累积是破坏性的，而解被误差淹没．不言而喻，这种算法被称为是**不稳定的**（unstable）．

1.15 解释浮点十进制数 $0.1066*10^4$.

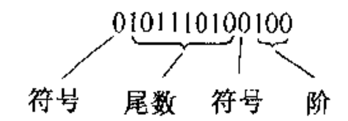
解 ☞ 显然,阶 4 使小数点右移 4 位,成为 1066. 类似,$0.1066*10^{-2}$是 0.001066.

1.16 解释浮点二进制符号 $+0.10111010*2^4$

解 ☞ 这里的阶将二进制的小数点向右移动 4 位,使之成为 1011.1010.它等价于十进制数 $11+\frac{5}{8}$或 11.625 类似地,$+0.10111010*2^{-1}$是 0.01011101,显然,它是原先所给定数的 1/32 倍.

1.17 解释浮点二进制记号 0101110100100.除了它们的符号,尾数使用 8 位而阶使用 3 位.

解 ☞ 第一和第十位上的零表示正号,



二进制的小数点取在尾数的前面,按照这些来理解,我们又一次得到了 $+0.10111010*2^4$.类似地,并以相同的约定,$+0.10111010*2^{-1}$变为 0101110101001,最后 4 位数字的含义是阶为 -1.

1.18 利用前题中的约定,将以下浮点数相加:

0101101110010
0100011001100

解 ☞ 不管怎么样,二进制小数点必须"对齐".对于记号的解释引出如下的和:

$$
\begin{array}{r}
10.110111 \\
+\ 0.000010001100 \\
\hline
=10.111001001100
\end{array}
$$

使用输入时的形式,它就变成

0101110010010

这里,除了符号,尾数仍取 8 位,阶取 3 位.由于机器的容量,最后 6 位二进制数字被删除,这就产生了舍入误差.

1.19 什么是溢出?

解 ☞ 仍利用我们虚构的机器中的约定,能表示的最大数是 0111111110111,这个数中,尾数和阶均是最大的.二进制小数点向右移动 7 位,等价于 1111111.1,这是十进制的数 $127+\frac{1}{2}$或 2^7-2^{-1}.在已知的约定下,任何大于这个数的数都不能被表示,而这就称为一个溢出(overflow).

1.20 什么是下溢?

解 ☞ 除了零和负数,在我们虚拟的机器中,形式上能被表示的最小数是 0000000011111.然而,鉴于各种理由,这样做是适当的:要求尾数的第一位数是 1,然后来确定阶.这就是有名的**规格化形式**(normalized form).零必须作为一个例外 如果规格化是必须的,则最小的正数变成 0100000001111.在十进制数中,它是 $2^{-1}*2^{-7}$或 2^{-8}.小于这个数的任何正数都不能用该机器来表示,而这就称为**下溢**(underflow).表示数字的任何浮点系统都会受到这种限制,并且都将使用溢出和下溢的概念.

1.21 假想有一个甚至更简单的浮点系统,在这个系统中尾数仅有 3 个二进制数字,且阶只是 -1,0 或 1,那么,这些数是如何分布在一条实线上的?

解 ☞ 假定规格化了,则这些数除了阶之外有如下形式:1××.于是,整个集合包含三个子集,每一个子集包括 4 个数如下:

0.0100	0.0101	0.0110	0.0111	(对于阶为 -1)
0.100	0.101	0.110	0.111	(对于阶为 0)
1.00	1.01	1.10	1.11	(对于阶为 1)

它们被标在图 1.2 中

注意:较小的数较稠密的存贮.从前一组到后一组,其间隔从$\frac{1}{16}$增加到$\frac{1}{4}$.显然,这是由于我们仅

有3位有效数字(第一位固定为1),它随着阶的增加提供逐级放大的量的缘故.例如,1.005在这里是不能表示的,它需要4位有效数字.而在这部分范围内,集合没有那么稠密.现实中的浮点系统处在更复杂的情况下,但具有相同的特征.并且,有效数字的思想与相对误差是有关的.

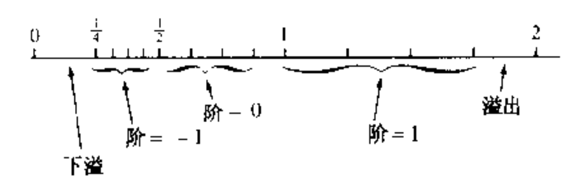


图 1.2

1.22 假定数 x 由浮点二进制符号表示,四舍五入为 n 位尾数,又假定它是规格化了的.那么,由舍入引起的绝对误差(absolute error)和相对误差的界是什么?

解 ☞ 舍入引起的误差至多为二进制数第 $n+1$ 位的一个单位,或第 n 位的半个单位.于是

$$\text{绝对误差} \leqslant 2^{-n-1}.$$

而对于相对误差,我们必须考虑其真值 x.规格化意味着其尾数不小于$\frac{1}{2}$,这导致如下的界:

$$|\text{相对误差}| \leqslant \frac{2^{-n-1}}{2^{-1}} = 2^{-n}.$$

设 $\mathrm{fl}(x)$表示 x 的浮点符号,用它来改写是有益的.于是,

$$\text{相对误差} = \frac{\mathrm{fl}(x)-x}{x} = E$$

或

$$\mathrm{fl}(x) = x(1+E) = x + xE,$$

其中$|E|\leqslant 2^{-n}$.舍入误差的运算可以看成由一个扰动值(perturbed value) $x+xE$ 去代替 x,而这个扰动相对来说是小的.

1.23 找出由两个浮点数相加造成的相对误差界.

解 ☞ 设这两个数是 $x=m_1*2^e$, $y=m_2*2^f$,而 y 较小.则 m_2 必须向右移动 $e-f$ 位(对齐二进制的小数点),然后尾数相加,将结果规格化并加以舍入.这样存在两种可能性:或者对二进制小数点左边发生溢出(这里所说的溢出不是题1.19意义上的溢出),或者不发生.第一种可能性由

$$1 \leqslant |m_1 + m_2 * 2^{f-e}| < 2$$

来描述,而第二种可能性由

$$\frac{1}{2} \leqslant |m_1 + m_2 * 2^{f-e}| < 1$$

来描述.若确实发生这种溢出,将要求向右移一位,然后我们有

$$\mathrm{fl}(x+y) = [(m_1 + m_2 * 2^{f-e}) * 2^{-1} + \epsilon] * 2^{e+1},$$

其中,ϵ 是舍入误差.这可以改写为

$$\mathrm{fl}(x+y) = (x+y)\left(1 + \frac{2\epsilon}{m_1 + m_2 * 2^{f-e}}\right)$$
$$= (x+y)(1+E),$$

而$|E|\leqslant 2\epsilon \leqslant 2^{-n}$.

若不发生这种溢出,则

$$\mathrm{fl}(x+y) = [(m_1 + m_2 * 2^{f-e}) + \epsilon] * 2^e$$
$$= (x+y)\left(1 + \frac{\epsilon}{m_1 + m_2 * 2^{f-e}}\right)$$
$$= (x+y)(1+E),$$

其中 E 的界同前.

浮点数减法的相应结果,将在题1.45中给出.

1.24 找出两个浮点数相乘的相对误差界.

解 ☞ 再一次设这两个数为 $x=m_1*2^e$ 和 $y=m_2*2^f$，则 $xy=m_1m_2*2^{e+f}$. 因为已规格化，故 $\frac{1}{4}\leqslant|m_1m_2|<1$. 这意味着规格化使乘积将至多左移一位. 因此，舍入后的积将或者是 $m_1+m_2+\epsilon$，或者是 $2m_1m_2+\epsilon$，而 $|\epsilon|\leqslant 2^{-n-1}$. 这能概述如下：

$$\mathrm{fl}(xy)=\begin{cases}(m_1m_2+\epsilon)*2^{e+f}, & 若\ |m_1m_2|\geqslant\frac{1}{2},\\ (2m_1m_2+\epsilon)*2^{e+f-1}, & 若\frac{1}{2}>|m_1m_2|\geqslant\frac{1}{4}\end{cases}$$

$$=m_1m_2*2^{e+f}\begin{cases}1+\frac{\epsilon}{m_1m_2}, & 若\ |m_1m_2|\geqslant\frac{1}{2},\\ 1+\frac{\epsilon}{2m_1m_2}, & 若\frac{1}{2}>|m_1m_2|\geqslant\frac{1}{4}\end{cases}$$

$$=xy(1+E),$$

而 $|E|\leqslant 2|\epsilon|\leqslant 2^{-n}$.

题 1.46 中，对于除法运算概述了一个类似结果. 这意味着在所有的 4 种算术运算中，利用浮点数，引来的相对误差不超过尾数的最末位有效数字上的 1.

1.25 估计利用浮点运算计算和

$$x_1+x_2+\cdots x_k$$

所产生的误差.

解 ☞ 我们考虑部分和 s_i. 设 $s_1=x_1$，则

$$s_2=\mathrm{fl}(s_1+x_2)=(s_1+x_2)(1+E_1).$$

正如题 1.23 中所证明的，E_1 的界为 2^{-n}. 改写

$$s_2=x_1(1+E_1)+x_2(1+E_1),$$

继续下去

$$s_3=\mathrm{fl}(s_2+x_3)=(s_2+x_3)(1+E_2)$$
$$=x_1(1+E_1)(1+E_2)+x_2(1+E_1)(1+E_2)+x_3(1+E_2),$$

而最后有

$$s_k=\mathrm{fl}(s_{k-1}+x_k)=(s_{k-1}+x_k)(1+E_{k-1})$$
$$=x_1(1+c_1)+x_2(1+c_2)+\cdots+x_k(1+c_k),$$

其中，对于 $i=2,\cdots,k$，

$$1+c_i=(1+E_{i-1})(1+E_i)\cdots(1+E_{k-1}),$$

且 $1+c_1=1+c_2$. 鉴于 E_j 的界相同，对于 $1+c_i$，现在我们有如下估计：

$$(1-2^{-n})^{k-i+1}\leqslant 1+c_i\leqslant(1+2^{-n})^{k-i+1},$$

故可概括出

$$\mathrm{fl}\left(\sum_{j=1}^{k}x_j\right)=\left(\sum_{j=1}^{k}x_j\right)(1+E),$$

其中，

$$E=\sum_{j=1}^{k}x_jc_j\Big/\sum_{j=1}^{k}x_j.$$

注意：若真和 $\sum x_j$ 相对于 x_j 是小的，则相对误差 E 可能是大的，这是由减法引起的相消结果，早在题 1.12 中就被注意到了.

1.26 阐述向前误差分析.

解 ☞ 假定 $A(B+C)$ 的值是由近似值 a,b,c 计算出来的，这些近似值误差量是 e_1,e_2,e_3，则真值

$$A(B+C)=(a+e_1)(b-e_2+c+e_3)=ab+ac+误差,$$

其中，

$$误差=a(e_2+e_3)+be_1+ce_1+e_1e_2+e_1e_3.$$

假设有统一的误差界 $|e_i|\leqslant e$，且误差之积可以忽略不计，那么，我们得到

$$|\text{误差}| \leqslant (2|a|+|b|+|c|)e.$$

这个典型的过程被称为向前误差分析(forward error analysis),原则上它能运用于任何算法.然而,通常这种分析不是不知所措的就是冗长乏味的.除此之外,所产生的这个界,通常是非常保守的,只适应于想知道最坏可能发生什么情况.在此例中,很少被留意但确实出现的一点是:看上去 a 的值比 b 和 c 的值敏感两倍.

1.27 什么是向后误差分析?

解 ☞ 向后误差分析(backward error analysis)的本质思想是:先接受计算结果,然后去确定能产生它的输入数据的范围.在这里,重要的是不要误解这样做的动机即不存在修改输入数据使之适应计算结果的企图.若完成了向后误差分析而显示出所获结果与输入数据的观测或舍入误差的范围相一致,则结果可以信赖.反之,则在另外的地方存在误差的主要来源,按推测,它存在于算法本身.

1.28 说明在题 1.23 中的误差分析就是向后误差分析.

解 ☞ 1.23 中得到的结果是

$$\text{fl}(x+y)=(x+y)(1+E),$$

$|E|\leqslant 2^{-n}$,其中,n 是二进制尾数的位数.将它改写为

$$\text{fl}(x+y)=x(1+E)+y(1+E).$$

回顾题 1.22,我们发现计算所得的和,即 $\text{fl}(x+y)$,仍是与原始数据 x 和 y 相差不大于一个舍入误差界 E 的两个数的真和,也就是说,输出可由恰当地落在认可的误差限制内的输入数据来解释.

1.29 说明在题 1.24 中所做的分析是向后误差分析.

解 ☞ 我们发现,

$$\text{fl}(xy)=xy(1+E)$$

可看成是 x 乘以 $y(1+E)$ 的积.这意味着计算出来的 $\text{fl}(xy)$ 也是与原来的 x,y 的差别不大于舍入误差的两个数的真积(true product),这与输入数据在我们认可的误差限制内是一致的.

1.30 在题 1.25 中,向后误差分析说明了什么?

解 ☞ 首先,方程

$$\text{fl}\left(\sum_{j=1}^{k}x_j\right)=x_1(1+c_1)+x_2(1+c_2)+\cdots+x_k(1+c_k)$$

表示 k 个数 x_1 到 x_k 的浮点运算之和也是 k 个数的真和,这 k 个数与原来的 k 个 x_j 差在相对误差 c_j 上.不幸的是,题 1.25 中得到的估计也表明这些误差可能远远大于单一的舍入误差.

1.31 取 L_2 范数,由先证 Cauchy-Schwarz 不等式

$$\left(\sum a_ib_i\right)^2\leqslant\left(\sum a_i^2\right)\left(\sum b_i^2\right),$$

来证明向量长度的三角性质.

解 ☞ 一种有趣的证明从注意到 $\sum(a_i-b_ix)^2$ 非负时开始,于是二次方程

$$\left(\sum b_i^2\right)x^2-2\left(\sum a_ib_i\right)x+\sum a_i^2=0$$

不能有不同的实根,这要求

$$4\left(\sum a_ib_i\right)^2-4\sum a_i^2\sum b_i^2\leqslant 0,$$

约去 4,我们便得到 Cauchy-Schwarz 不等式.

现在,仅用一点点代数知识,就能立即得到三角不等式,写成分量形式,有

$$[(v_1+w_1)^2+\cdots+(v_n+w_n)^2]^{\frac{1}{2}}\leqslant(v_1^2+\cdots+v_n^2)^{\frac{1}{2}}+(w_1^2+\cdots+w_n^2)^{\frac{1}{2}}.$$

对它平方,合并同类项,再平方,利用 Cauchy-Schwarz 不等式就将得到所需的结果(见题 1.50).

1.32 试证,当 p 趋向于无穷,向量的 L_p 范数逼近于 $\max|v_i|$.

证 ☞ 假设 v_m 是绝对值最大的分量,并将和改写为

$$|v_m|\left(1+\sum_{i\neq m}\left|\frac{v_i}{v_m}\right|^p\right)^{1/p},$$

括号内除了第一项外,所有的项趋于零,于是有所需的结果.

1.33 试证,对于单位向量 V,定义的 $\|A\| = \max\|AV\|$ 能满足引言中所给的性质 1～3.

证 ☞ 这些性质很容易从相伴向量范数对应的性质得到.由于 AV 是向量,$\|AV\| \geqslant 0$,故 $\|A\| \geqslant 0$.若 $\|A\| = 0$,而 A 哪怕仅有一个不为零的元素,则能选择某 V 使得 AV 的一个分量是正的,这与 $\|AV\| = 0$ 矛盾.这就证出了性质 1.

其次,我们有

$$\|cA\| = \max\|cAV\| = \max|c|\cdot\|AV\| = |c|\cdot\|A\|.$$

这证明了性质 2.性质 3 可类似地处理.

1.34 单位矩阵(identity matrix)的 L_1, L_2 和 L_∞ 范数分别是什么?

解 ☞ 它们均为 1.因为 V 是单位向量,我们有

$$\|I\| = \max\|IV\| = \max\|V\| = 1.$$

1.35 矩阵 $\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$ 的 L_1, L_2 和 L_∞ 范数分别是什么?

解 ☞ 我们有

$$AV = \begin{bmatrix}v_1 + v_2\\v_1 + v_2\end{bmatrix}.$$

为了简单起见,设 v_1, v_2 非负.由于 V 是 L_1 范数中的单位向量,因而对于 L_1 范数,我们相加有 $\|AV\| = 2(v_1 + v_2) = 2$,于是 $\|A\|_1 = 2$.对于 L_2 范数,我们需对两分量平方然后相加,得到 $2(v_1^2 + 2v_1v_2 + v_2^2)$.在这种范数中,$v_1^2 + v_2^2 = 1$,于是我们对 v_1v_2 取最大值,由基本的微积分运算可得 $v_1 = v_2 = \frac{1}{\sqrt{2}}$,立得 $\|A\|_2 = 2$.最后,因为使用 L_∞ 范数我们要找的是最大分量,所以 $\|AV\|_\infty = v_1 + v_2$.由于使用这种范数,$v_1, v_2$ 均不超过 1,故这里最大值仍是 2.利用下面的题或它的相伴题的结果,L_1 和 L_∞ 范数能很快被解出.

1.36 试证

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|.$$

证 ☞ 选择一个所有分量为 1 且符号与 a_{ij} 相匹配的向量 V,使得 $\sum |a_{ij}|$ 最大.于是,$\sum a_{ij}v_j$ 是 AV 的一个等于该最大值的元素,并且,它显然不可能被超过.由于 V 的范数是 1,故 A 的范数仍取此值.L_1 范数的类似结果留作题 1.52

1.37 证明 $\|AV\| \leqslant \|A\|\cdot\|V\|$.

证 ☞ 根据 $\|A\|$ 的定义,对于单位向量 U,我们有

$$\|AU\| \leqslant \max_U \|AU\| = \|A\|,$$

选取 $U = V/\|V\|$ 并应用性质 2,

$$\left\|A\left(\frac{V}{\|V\|}\right)\right\| \leqslant \|A\|,\quad \|AV\| \leqslant \|A\|\,\|V\|.$$

1.38 证明 $\|AB\| \leqslant \|A\|\cdot\|B\|$.

证 ☞ 我们重复利用题 1.37 的结果,

$$\begin{aligned}\|AB\| &= \max\|ABU\| \leqslant \max\|A\|\cdot\|BU\| \leqslant \max\|A\|\cdot\|B\|\cdot\|U\|\\ &= \|A\|\cdot\|B\|.\end{aligned}$$

补 充 题

1.39 利用支撑理论

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots$$

计算 $\frac{1}{0.982}$,其中 $x = 0.018$.

1.40 当误差不超过 0.005 时，数字准确到 2 位小数，下面的平方根取自一个表.

n	11	12	13	14	15	16	17	18	19	20
$\sqrt{n}$四舍五入到 3 位小数	3.317	3.464	3.606	3.742	3.873	4.000	4.123	4.243	4.359	4.472
$\sqrt{n}$四舍五入到 2 位小数	3.32	3.46								
舍入误差的近似值	−0.003	−0.004								

试将每个数舍入到两位小数并注出舍入误差的数量大小. 这些舍入误差与误差最大值 0.005 相比情况如何？理论上这 10 个数总舍入误差在 10(−0.005)到 10(0.005)之间，而实际上是多少？它与“概率误差”$\sqrt{10}(0.005)$相比，情况如何？

1.41 假设 N 个数均准确到给定的位数，求其和. 用概率误差公式来估计，当 N 大约多大时，计算出来的和的最后一位数字将可能无意义？当 N 大约多大时，和的最后两位数可能无意义？

1.42 序列 $J_0, J_1, J_2, \cdots$ 由

$$J_{n+1} = 2nJ_n - J_{n-1}$$

定义，$J_0 = 0.765198$，$J_1 = 0.440051$ 均准确到 6 位小数. 试计算 $J_2, \cdots, J_7$ 且将它们与下表中的准确值进行比较(这些准确值是由另一完全不同的方法得到的. 对误差的解释见下题)

n	2	3	4	5	6	7
精确的 J_n	0.114903	0.019563	0.002477	0.000250	0.000021	0.000002

1.43 试证：对于上题的序列，精确地有

$$J_7 = 36767J_1 - 21144J_0.$$

由给定的 J_0 和 J_1 值来计算，将得到同样不准确的值. 大系数乘给定的 J_0 和 J_1 值中的舍入误差，则合并后的结果含有一个大误差.

1.44 数 J_3 直到 6 位都将为零，按题 1.42 中的公式，实际得出什么？

1.45 试证浮点数减法所产生的误差以 2^{-n} 为界. 如在题 1.23 中那样，设 $x = m_1 * 2^e$，$y = m_2 * m^f$，则 $x - y = (m_1 - m_2 * 2^{f-e})2^e$，除非它为零，否则就有

$$2^{-n} \leqslant | m_1 - m_2 * 2^{f-e} | < 2.$$

对这个新的尾数进行规格化处理，也许小数点需要左移 $n-1$ 位，而实际的数 s 由

$$2^{-s-1} \leqslant | m_1 - m_2 * 2^{-ef} | < 2^{-s}$$

所决定. 今试证

$$\mathrm{fl}(x - y) = [(m_1 - m_2 * 2^{f-e}) * 2^s + \varepsilon] * 2^{e-s},$$

而最终

$$\mathrm{fl}(x - y) = (x - y)(1 + E),$$

其中 $|E| \leqslant 2^{-n}$.

1.46 试证浮点数除法所产生的误差以 2^{-n} 为界. 按照题 1.24 中的约定，让分子尾数的一半除以分母的尾数(这是为了避免商大于 1)而阶相减，它给出了

$$\frac{x}{y} = \left(\frac{m_1}{2m_2}\right) * 2^{e-f+1}$$

其中 $\frac{1}{4} \leqslant |m_1/2m_2| < 1$. 现仿效对乘法运算所作的分析步骤，再一次证明相对误差 E 如所述，是有界的.

1.47 分析内积计算

$$s_k = \mathrm{fl}(x_1y_1 + x_2y_2 + \cdots + x_ky_k).$$

它酷似题 1.25. 设

$$t_i = \mathrm{fl}(x_iy_i), \quad i = 1, \cdots, k,$$

接着令

$$s_1 = t_1 \quad s_i = \mathrm{fl}(s_{i-1} + t_i), \quad i = 1, \cdots, k$$

它造出了所求的内积 s_k. 今求出类似于那些在前面的题中得到的关系式和估计.

1.48 使用题 1.17 的约定,解释浮点符号 0100110011010(这是仅以 8 位尾数对0.1492最可能的接近).

1.49 仿效题 1.21,想象一个浮点系统,在该系统中规格化的尾数有 4 位,而阶为 $-1,0,1$. 证明这些数形成了三组,各有 8 个数;对应于它们的阶,一组落在 1/4 到 1/2 区间中,另一组在 1/2 到 1 区间中,而第三组在 1 与 2 之间.哪个正数会造成溢出? 哪个下溢?

1.50 完成在题 1.31 中开始的证明.

1.51 通过证明两个矩阵和的范数不超过它们范数的和来完成题 1.33.

1.52 通过对单位向量的适当选择(一个分量为 1,其余的为零),证明矩阵 A 的 L_1 范数可以从绝对值元素的最大列和算得,并与题 1.36 中相关的证明相比较.

1.53 证明对矩阵 $A=\begin{bmatrix} a & b \\ b & a \end{bmatrix}$, L_1, L_2 及 L_3 范数均相等.

1.54 证明对矩阵 $A=\begin{bmatrix} a & b \\ b & -a \end{bmatrix}$,其 L_2 范数为 $(a^2+b^2)^{1/2}$.

1.55 证明对矩阵 $A=\begin{bmatrix} a & a \\ a & b \end{bmatrix}$,可以得到一个使 $\|AV\|_2$ 为极大的向量 V,其形式为 $(\cos t, \sin t)^T$. 其中,在 $b^2=a^2$ 的情况下, $\cos 2t=0$,而在另外的情况下 $\tan 2t=2a/(a-b)$.

1.56 下面的信息已经被建议作为本行星生活着智慧生命的信号传播到外层空间.这里的想法是,无论在什么地方的任何形式的智慧生命都会理解这信息的智慧内涵,并由此推知这里存在着我们所拥有的智慧.该信息

11.00100100001111110110

的意义是什么?

1.57 若以 x, y 为分量的向量 $\mathbf{V}$ 表示平面上的一个点(x,y),则对应于取 L_2 范数的单位向量的点形成古典的单位圆.如图 1.1 所示,对 L_1 及 L_∞ 范数,该"圆"取作正方形.在一个有正方形街区的城市中,对出租车行进来说,哪一种是合适的范数(从一个交叉点出发,在给定距离中,找出所有的交叉点).在一个棋盘上,为什么对国王的行进而言,合适的范数是 L_∞ 范数?

第二章　配置多项式

多项式逼近

多项式逼近是数值分析中最古老的思想之一，而且是迄今仍最受重用的方法之一.对于一个函数 $y(x)$，用一个多项式 $p(x)$ 代替它有着众多的理由，其中最重要的也许是因为多项式便于计算，它仅涉及到简单的整数幂.它们的导数和积分也不难得到，并且仍然是多项式.多项式较之其他函数易于求根，故而流行用多项式代替其他函数是不难理解的.

逼近准则

差 $y(x)-p(x)$ 是逼近误差.显然，其核心思想是保持该误差合理地小.由于多项式简单，允许以不同的方法接近这个目标.在这些方法中我们考虑的是

1. 配置(collocation)
2. 密切(osculation)
3. 最小二乘(least squares)
4. 极小-极大(min-max)

配置多项式

配置多项式(collocation polynomial)是这一章及下面少数几章中的研究对象，在某些指定的点上，它与 $y(x)$ 重合，这种多项式与一般的多项式的许多性质在展开过程中起作用.

1. **存在和惟一性定理**指出，对于自变量 $x_0,\cdots,x_n$，恰好存在一个 n 次配置多项式，即，使得对这些变量，$y(x)=p(x)$.存在性将由实际展示在后继的章节中的多项式证实.惟一性将在本章中被证明.它是多项式某些基本性质的一个结果.
2. **辗转相除法**.任何多项式可以表示为
$$p(x)=(x-r)q(x)+R,$$
其中，r 是任意数，$q(x)$ 是一个 $n-1$ 次多项式，而 R 是一个常数.它有两个直接的推论.
3. **剩余定理**.(remainder theorem)指出，$p(r)=R$.
4. **因式定理**.(factor-theorem)指出，若 $p(r)=0$，则 $x-r$ 是 $p(x)$ 的因式.
5. **零点限制**.一个 n 次多项式至多有 n 个零点，这意味着方程 $p(x)=0$ 至多有 n 个根.作为需要证明的惟一性定理，是一个直接的推论.正如将证明的那样.
6. **综合除法(synthetic division)**.对于获取 $q(x)$ 和 R 的辗转相除法来说，是一个经济的程序(或算法).通常它被用于求 R，由剩余定理，$R=p(r)$.求 $p(r)$ 的这条路也许比直接计算多项式的值更好.
7. **乘积**.$\pi(x)=(x-x_0)(x-x_1)\cdots(x-x_n)$ 在配置理论中起着重要作用.注意，在配置自变量 $x_0,x_1,\cdots,x_n$ 处，乘积为零.配置多项式的误差将被证明是
$$y(x)-p(x)=\frac{y^{(n+1)}(\xi)\pi(x)}{(n+1)!},$$
其中，ξ 取决于 x.并且，倘若 x 是配置端点，则 ξ 位于端点间.注意到这个公式在 $x_0,x_1,\cdots,x_n$ 处为零，从而在这些点上，$p(x)$ 确实与 $y(x)$ 相配置，而在其他的地方，我们将 $p(x)$ 看作是对 $y(x)$ 的逼近.

题 解

2.1 证明:任一多项式 $p(x)$ 能表示为

$$p(x) = (x - r)q(x) + R,$$

其中,r 是任意数,$q(x)$ 是一个 $n-1$ 次多项式,而 R 是一个常数.

证 ☞ 这是辗转相除法的一个例子.设 $p(x)$ 为 n 次多项式,

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0,$$

则

$$p(x) - (x - r)a_n x^{n-1} = q_1(x) = b_{n-1}x^{n-1} + \cdots$$

将不大于 $n-1$ 次.类似地有,

$$q_1(x) - (x - r)b_{n-1}x^{n-2} = q_2(x) = c_{n-2}x^{n-2} + \cdots$$

将不大于 $n-2$ 次.依次类推,最终我们可得到一个 0 次多项式 $q_n(x)$,即一个常数,记这个常数为 R,我们有

$$p(x) = (x - r)(a_n x^{n-1} + b_{n-1}x^{n-2} + \cdots) + R = (x - r)q(x) + R.$$

2.2 证明 $p(r) = R$.而这被称为剩余定理.

证 ☞ 设题 2.1 中 $x = r$,立得 $p(r) = 0 \cdot q(r) + R$.

2.3 利用 $r = 2$, $p(x) = x^3 - 3x^2 + 5x + 7$,来阐述"综合除法",用以完成题 2.1 中所述的除法.

解 ☞ "综合除法"只不过是题 2.1 中相同运算的缩写形式,它仅出现各系数.对于上述的 $p(x)$ 和 r,开始的格式是

$r=2$	1	−3	5	7 ← $p(x)$ 的系数
	1			

"乘以 r 且相加"三次后,完成了格式.

$r=2$	1	−3	5	7
		2	−2	6
	1	−1	3	13 ← 此为数 R

（1 −1 3 为 $q(x)$ 的系数）

于是,$q(x) = x^2 - x + 3$, $R = f(2) = 13$.由计算 $(x - r)q(x) + R$ 可验证,这就是 $p(x)$.对于寻找 $q(x)$ 而言,"长除法(long division)"也是有用的.它从常见的格式

$$(x - 2)\ \sqrt{x^3 - 3x^2 + 5x + 7}$$

开始.将产生此结果的计算与刚才完成的"综合"除法相比较,易见两者是等价的.

2.4 证明:若 $p(r) = 0$,则 $x - r$ 是 $p(x)$ 的一个因式.这就是**因式定理**,那剩下的因式为 $n-1$ 次.

证 ☞ 若 $p(r) = 0$,则 $0 = 0 \cdot q(x) = R$,从而 $R = 0$.于是,

$$p(x) = (x - r)q(x).$$

2.5 证明 n 次多项式至多有 n 个零点,这意味着 $p(x) = 0$ 至多有 n 个根.

证 ☞ 假设存在 n 个根,记为 $r_1, r_2, \cdots, r_n$,将因式定理应用 n 次,则有

$$p(x) = A(x - r_1)(x - r_2)\cdots(x - r_n)$$

其中,A 有零次幂,是一个常数.这清楚地表明不可能存在其他的根(同时指出 $A = a_n$).

2.6 证明:至多有一个 n 次多项式在给定的自变量 x_k 处能取到指定的值 y_k,其中 $k = 0, 1, \cdots, n$.

证 ☞ 假设存在两个这样的多项式 $p_1(x)$ 和 $p_2(x)$,那么其差 $p(x) = p_1(x) - p_2(x)$ 将不大于 n

次,并在所有的自变量 x_k 处 $p(x_k)=0$.由于存在 $n+1$ 个这种自变量使多项式为零与上述问题矛盾,故至多有一个 n 次多项式能取到这些指定的值.以下章节以许多实用的形式来展示这种多项式.它被称为**配置多项式**.

2.7 假设一个 n 次多项式 $p(x)$ 与函数 $y(x)$ 在 $x=x_0,x_1,\cdots,x_n$ 处取相同的值(这被称为两个函数的配置,而 $p(x)$ 是配置多项式),求 $p(x)$ 与 $y(x)$ 之差的公式.

解 ☞ 由于在配置点上,两者的差是零,我们可预料有形如

$$y(x)-p(x)=C(x-x_0)(x-x_1)\cdots(x-x_n)=C\pi(x)$$

的结果.这可取作 C 的定义.今考虑如下函数 $F(x)$:

$$F(x)=y(x)-p(x)-C\pi(x).$$

对于 $x=x_0,x_1,\cdots,x_n$, $F(x)=0$.而若我们选一个新的自变量 x_{n+1},以及

$$C=\frac{y(x_{n+1})-p(x_{n+1})}{\pi(x_{n+1})},$$

则 $F(x_{n+1})$ 也将等于零,此时 $F(x)$ 至少有 $n+2$ 个零点.那么由 Rolle 定理,$F'(x)$ 定有 $n+1$ 个零点在 $F(x)$ 的零点之间;而 $F''(x)$ 定有 n 个零点在 $F'(x)$ 的零点之间.依此连续应用 Rolle 定理,最终证出,在 x_0 到 x_n 区间,$F^{(n+1)}(x)$ 至少存在一个零点,比如说,在 $x=\xi$ 处.现计算此导数:想到 $p(x)$ 的 $n+1$ 阶导数是零,令 $x=\xi$,有

$$0=y^{(n+1)}(\xi)-C(n+1)!,$$

这可定出 C,代入前面的式子有

$$y(x_{n+1})-p(x_{n+1})=\frac{y^{(n+1)}(\xi)}{(n+1)!}\pi(x_{n+1}).$$

由于 x_{n+1} 可以是 x_0 与 x_n 之间除 $x_0,x_1,\cdots,x_n$ 之外的任一自变量,并因为我们的结果对 $x_0,x_1,\cdots,x_n$ 也显然成立,我们用无下标的 x 来代替 x_{n+1},有

$$y(x)-p(x)=\frac{y^{(n+1)}(\xi)}{(n+1)!}\pi(x).$$

尽管数 ξ 通常是不能确定的,但这个结果通常仍相当有用,因为我们能不依赖 ξ 去估计 $y^{(n+1)}(\xi)$.

2.8 找出一个取值 $y(0)=1$ 且 $y(1)=0$,或取值如表

x_k	0	1
y_k	1	0

的一次多项式.

解 ☞ 根据验算或初等几何学,立即有所需结果 $p(x)=1-x$.这是只提供了贫乏数据的一个配置多项式.

2.9 函数 $y(x)=\cos\frac{1}{2}\pi x$ 仍取题 2.8 的指定值.求其差 $y(x)-p(x)$.

解 ☞ 根据题 2.7,当 $n=1$ 时,

$$y(x)-p(x)=-\frac{\pi^2\cos\frac{1}{2}\pi\xi}{8}x(x-1).$$

即使没有定出 ξ,我们也能根据

$$|y(x)-p(x)|\leqslant\frac{\pi^2}{8}x(x-1)$$

估计出这个差.将 $p(x)$ 看作 $y(x)$ 的一个线性逼近,其误差估计是简单的,不过过估了.在 $x=\frac{1}{2}$ 时,它指出的误差大致为 0.3,而实际的误差 $\cos\frac{1}{4}\pi-\left(1-\frac{1}{2}\right)\approx0.2$.

2.10 当次数 n 无限增加时,所得的配置多项式序列是否收敛于 $y(x)$?

解 ☞ 答案比较复杂.正如后面将出现的,若仔细选择配置点 x_k 及合理的函数 $y(x)$,则收敛是肯定的.但对于最一般的点 x_k 等距的情况,可能会发散.对某些 $y(x)$,多项式序列对所有自变量 x 均

是收敛的，而对另一些函数，收敛仅限于一个有限区间，而误差 $y(x)-p(x)$ 的振荡形式如图 2.1 所示.在收敛区域内，振荡消失而 $\lim(y-p)=0$.但在那个区间外，$y(x)-p(x)$ 随着 n 的增大而任意增大.该振荡产生于 $\pi(x)$ 因式，其幅度受 $y(x)$ 的导数的影响，这个误差特性严重地限制了对高阶配置多项式的使用.

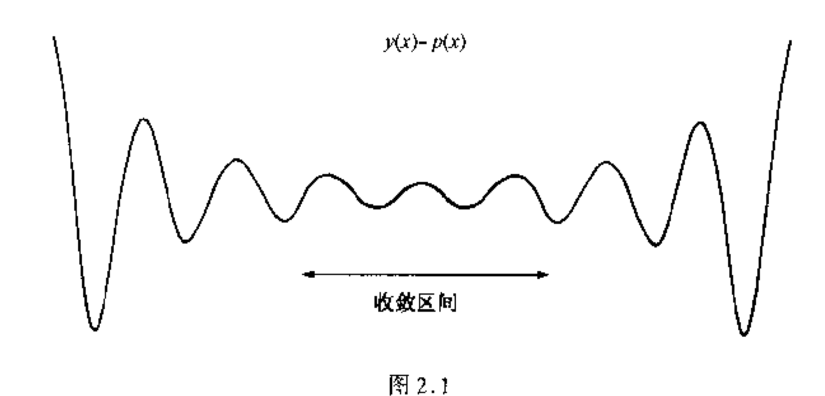


图 2.1

补　充　题

2.11 应用综合除法，以 $x-1$ 除 $p(x)=x^3-x^2+x-1$.注意：$R=p(1)=0$，从而 $x-1$ 是 $p(x)$ 的一个因式且 $r=1$ 是 $p(x)$ 的一个零点.

2.12 对 $p(x)=2x^4-24x^3+100x^2-168x+93$ 应用综合除法，计算 $p(1)$（除以 $x-1$ 且取余数 R）.同时计算 $p(2)$，$p(3)$，$p(4)$ 及 $p(5)$.

2.13 为了找出取如下值

x_k	0	1	2
y_k	0	1	0

的一个二次多项式，我们可写出 $p(x)=A+Bx+Cx^2$，代入后获得条件

$$0=A,\quad 1=A+B+C,\quad 0=A+2B+4C,$$

解出 A，B，C，从而定出这个配置多项式.理论上，对于高次多项式，可用相同的方法得到，但是，更有效的算法将会被开发.

2.14 设函数 $f(x)=\sin\frac{1}{2}\pi x$ 仍取题 2.13 中的指定值.试用题 2.7 证明

$$y(x)-p(x)=\frac{\pi^3\cos\frac{1}{2}\pi\xi}{48}x(x-1)(x-2),$$

其中，ξ 依赖于 x.

2.15 继题 2.14，试证

$$|y(x)-p(x)|\leqslant\left|\frac{\pi^3}{48}x(x-1)(x-2)\right|.$$

它用来估计以配置多项式 $p(x)$ 作为 $y(x)$ 的一个逼近的精度.当 $x=\frac{1}{2}$ 时计算这个估计误差，并与实际误差相比较.

2.16 对于 $x=\frac{1}{2}$，比较 $y'(x)$ 与 $p'(x)$.

2.17 对于 $x=\frac{1}{2}$，比较 $y''(x)$ 与 $p''(x)$.

2.18 比较 $y(x)$ 与 $p(x)$ 在区间 $(0,2)$ 上的积分.

2.19 找出取值如下表

x_k	0	1	2	3
y_k	0	1	16	81

的惟一的三次多项式 $p(x)$.

2.20　设函数 $y(x)=x^4$,也取上题中的给定值.对于差 $y(x)-p(x)$,利用题 2.7 写出一个公式.

2.21　在区间(0,3)上,$|y(x)-p(x)|$的最大值是多少?

第三章　有限差分

有限差分

几个世纪以来，有限差分(finite difference)对数学家们具有强烈的吸引力. Isaac Newtow 是它的一个特别重要的使用者，而且许多课题还起源于他. 给出一个离散的函数，即给出一个自变量 x_k 的集合，其中，每个 x_k 对应一个 y_k，又假设这些自变量是等距的，于是 $x_{k+1} - x_k = h$，相应的 y_k 值的差被记为

$$\Delta y_k = y_{k+1} - y_k,$$

并被称为一阶差分(first difference). 这些一阶差分的差被记为

$$\Delta^2 y_k = \Delta(\Delta y_k) = \Delta y_{k+1} - \Delta y_k = y_{k+2} - 2y_{k+1} + y_k.$$

被称为二阶差分(second difference). 一般

$$\Delta^n y_k = \Delta^{n-1} y_{k+1} - \Delta^{n-1} y_k,$$

定义为 n 阶差分(nth difference).

如下的差分表(differences table)是展示有限差分的一个标准格式. 除了 x_k，y_k，它的对角线模式使得每一个表值成为它左边两相邻者的差：

x_0	y_0				
		Δy_0			
x_1	y_1		$\Delta^2 y_0$		
		Δy_1		$\Delta^3 y_0$	
x_2	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$
		Δy_2		$\Delta^3 y_1$	
x_3	y_3		$\Delta^2 y_2$		
		Δy_3			
x_4	y_4				

每一个差分都能被证明是第二列中 y 值的一个组合. 一个简单的例子是 $\Delta^3 y_0 = y_3 - 3y_2 + 3y_1 - y_0$；其一般结果是

$$\Delta^k y_0 = \sum_{i=0}^{k} (-1)^i \binom{k}{i} y_{k-i},$$

其中，$\binom{k}{i}$ 是二项式的一个系数.

差分公式

就初等函数而言，差分公式(differencc formula)与微积分公式有些平行. 例子如下：

1. 一个**常数函数的差分**是零，记为

$$\Delta C = 0,$$

　其中，C 表示一个常数(与 k 无关).

2. 对于**一个常数乘另一函数**，我们有

$$\Delta(Cu_k) = C\Delta u_k.$$

3. 两个函数之**和的差分**是它们的差分之和：

$$\Delta(u_k + v_k) = \Delta u_k + \Delta v_k.$$

4. **线性性质**归纳了上面的两个结果：

$$\Delta(C_1u_k + C_2v_k) = C_1\Delta u_k + C_2\Delta v_k,$$

其中，C_1，C_2 是常数.

5. **积的差分**由公式

$$\Delta(u_kv_k) = u_k\Delta v_k + v_{k+1}\Delta u_k$$

给出.在此应注意变量 $k+1$.

6. **商的差分**是

$$\Delta\left(\frac{u_k}{v_k}\right) = \frac{v_k\Delta u_k - u_k\Delta v_k}{v_{k+1}v_k}$$

再一次提请注意变量 $k+1$.

7. **幂函数的差分**由

$$\Delta C^k = C^k(C-1)$$

给出.特别，$C=2$ 有 $\Delta y_k = y_k$

8. **正弦与余弦函数的差分**也令人回顾微积分中相应的结果，但细节却颇缺乏那种魅力：

$$\Delta(\sin k) = 2\sin\frac{1}{2}\cos\left(k+\frac{1}{2}\right),$$

$$\Delta(\cos k) = -2\sin\frac{1}{2}\sin\left(k+\frac{1}{2}\right).$$

9. **对数函数的差分**同样令人失望，当 $x_k = x_0 + kh$，我们有

$$\Delta(\log x_k) = \log\left(1+\frac{h}{x_k}\right).$$

当 h/x_k 非常小时，$\Delta(\log x_k)$近似为 h/x_k，而在对数的微分运算中，它是 x 的倒数，非常显著，两者相距甚远.

10. **单位误差函数**(unit error function)，对于这种在一个单一的点上 $y_k=1$，而在其他点上是零的函数，有一个带交错符号的逐次的二项式系数构成的差分表.在一个 y_k 值的表中孤立误差的检测可基于单位误差函数的这个性质.

11. **振荡误差函数**(oscillating error function)，对于这种交替有 $y_k=\pm1$ 的函数，有一个带交错符号的、由 2 的逐次幂组成的差分表.

12. **其他特别重要的函数**在后继的章节中将被研究，而差分与微分运算两者间的关系将是一件继续关心的事.

题　　解

3.1　根据表 3.1 中 x_k，y_k 两列所展示的离散函数，计算到三阶差分(为方便计，整数变量 k 也列于表中).

表 3.1

k	x_k	y_k	Δy_k	$\Delta^2 y_k$	$\Delta^3 y_k$
0	1	1			
			7		
1	2	8		12	
			19		6
2	3	27		18	
			37		6
3	4	64		24	
			61		6
4	5	125		30	
			91		6
5	6	216		36	
			127		6
6	7	343		42	
			169		
7	8	512			

解 ☞ 所要求的差分在剩下的三列中.表 3.1 被称为差分表,它的对角线结构已成为展示差分的标准格式.差分列中的每一个表值是左边最近的两相邻者的差分.

任何这种表,展示差分如表 3.2 所示:

表 3.2

0	x_0	y_0			
			Δy_0		
1	x_1	y_1		$\Delta^2 y_0$	
			Δy_1		$\Delta^3 y_0$
2	x_2	y_2		$\Delta^2 y_1$	
			Δy_2		$\Delta^3 y_1$
3	x_3	y_3		$\Delta^2 y_2$	
			Δy_3		
4	x_4	y_4			

例如

$$\Delta y_0 = y_1 - y_0 = 8 - 1 = 7,$$
$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0 = 19 - 7 = 12,$$

等等.

3.2 题 3.1 中的函数的所有 4 阶和更高阶的差分是什么?

解 ☞ 任一 4 阶及以上差分是零.这是可立即获得结果的特殊情况.

3.3 证明 $\Delta^3 y_0 = y_3 - 3y_2 + 3y_1 - y_0$.

证 ☞ 根据表 3.2 或前面所给的差分定义,

$$\Delta^3 y_0 = \Delta^2 y_1 - \Delta^2 y_0 = (y_3 - 2y_2 + y_1) - (y_2 - 2y_1 + y_0)$$
$$= y_3 - 3y_2 + 3y_1 - y_0.$$

3.4 证明 $\Delta^4 y_0 = y_4 - 4y_3 + 6y_2 - 4y_1 + y_0$.

证 ☞ 根据定义,$\Delta^4 y_0 = \Delta^3 y_1 - \Delta^3 y_0$.利用题 3.3 的结果以及由提升所有的下标得到的几乎完全相同的式子

$$\Delta^3 y_1 = y_4 - 3y_3 + 3y_2 - y_1,$$

立即可得所需的结果.

3.5 证明:对于任意正整数 k,

$$\Delta^k y_0 = \sum_{i=0}^{k} (-1)^i \binom{k}{i} y_{k-i},$$

其中,

$$\binom{k}{i} = \frac{k!}{i!(k-i)!} = \frac{k(k-1)\cdots(k-i+1)}{i!}$$

正是熟知的二项式系数.

证 ☞ 用归纳法来证.对于 $k = 1,2,3,4$,结果已经成立,当 $k=1$ 时,根据的是定义.现假定当 k 是某个特定的整数 p 时,有

$$\Delta^p y_0 = \sum_{i=0}^{p} (-1)^i \binom{p}{i} y_{p-i},$$

提升所有的下标,我们又有

$$\Delta^p y_1 = \sum_{i=0}^{p} (-1)^i \binom{p}{i} y_{p-i+1}.$$

更改和式的指标,即令 $i = j + 1$,有

$$\Delta^p y_1 = y_{p+1} - \sum_{j=0}^{p-1}(-1)^j\binom{p}{j+1}y_{p-j}.$$

对和式的指标作一名称上的改变也是方便的.在我们的另一和式中,令 $i = j$,有

$$\Delta^p y_0 = \sum_{j=0}^{p-1}(-1)^j\binom{p}{j}y_{p-j} + (-1)^p y_0,$$

于是,

$$\Delta^{p+1}y_0 = \Delta^p y_1 - \Delta^p y_0 = y_{p+1} - \sum_{j=0}^{p-1}(-1)^j\left[\binom{p}{j+1} + \binom{p}{j}\right]y_{p-j} - (-1)^p y_0,$$

今利用

$$\binom{p}{j+1} + \binom{p}{j} = \binom{p+1}{j+1}$$

(见题 4.5)并作最后一个和标的变化,即令 $j + 1 = i$,有

$$\Delta^{p+1}y_0 = y_{p+1} + \sum_{i=1}^{p}(-1)^i\binom{p+1}{i}y_{p+1-i} - (-1)^p y_0$$

$$= \sum_{i=0}^{p+1}(-1)^i\binom{p+1}{i}y_{p+1-i}.$$

于是,当 $k = p + 1$ 时,结果成立,这完成了归纳法.

3.6 证明:对常数函数而言,所有的差分都是零.

证 ☞ 设对所有的 k, $y_k = C$,这是常数函数.于是对所有的 k,

$$\Delta y_k = y_{k+1} - y_k = C - C = 0.$$

3.7 证明 $\Delta(Cy_k) = C\Delta y_k$.

证 ☞ 这类似于微积分中的结果:$\Delta(Cy_k) = Cy_{k+1} - Cy_k = C\Delta y_k$.

本质上,这个问题包含了对相同变量 x_k 定义的两个函数.一个函数有值 y_k,另一个有值 $z_k = Cy_k$.我们已证出 $\Delta z_k = C\Delta y_k$.

3.8 考虑定义在相同自变量 x_k 集合上的两个函数.令这两个函数的值分别为 u_k 和 v_k.同样,考虑值

$$w_k = C_1 u_k + C_2 v_k$$

的第三个函数,其中 C_1, C_2 是常数(与 x 无关).证明:

$$\Delta w_k = C_1\Delta u_k + C_2\Delta v_k.$$

这是差分运算的**线性性质**.

证 ☞ 直接由定义来证明:

$$\Delta w_k = w_{k+1} - w_k = (C_1 u_{k+1} + C_2 v_{k+1}) - (C_1 u_k + C_2 v_k)$$

$$= C_1(u_{k+1} - u_k) + C_2(v_{k+1} - v_k) = C_1\Delta u_k + C_2\Delta v_k.$$

显然,相同的证明将适用于任何有限长度的和式.

3.9 利用题 3.8 中相同的符号,考虑具有值 $z_k = u_k v_k$ 的函数,证明 $\Delta z_k = u_k\Delta v_k + v_{k+1}\Delta u_k$.

证 ☞ 仍从定义开始:

$$\Delta z_k = u_{k+1}v_{k+1} - u_k v_k = u_{k+1}v_{k+1} - u_k v_{k+1} + u_k v_{k+1} - u_k v_k$$

$$= v_{k+1}(u_{k+1} - u_k) + u_k(v_{k+1} - v_k)$$

$$= u_k\Delta v_k + v_{k+1}\Delta u_k.$$

还可证明这样的结果:

$$\Delta z_k = u_{k+1}\Delta v_k + v_k\Delta u_k.$$

3.10 计算表 3.3 中前二列所展示的函数的差分.如果除了单独的 1 是一个单位误差之外,其他所有的值是零,则可视之为一个"误差函数"(error function)型.这个单位误差将如何影响各种差分?

解 ☞ 表 3.3 中的其余各列表示某些所需的差分.

表 3.3

x_0	0				
		0			
x_1	0		0		
		0		0	
x_2	0		0		1
		0		1	
x_3	0		1		-4
		1		-3	
x_4	1		-2		6
		-1		3	
x_5	0		1		-4
		0		-1	
x_6	0		0		1
		0		0	
x_7	0		0		
		0			
x_8	0				

这个误差影响到差分表的一个三角形部分.对于较高阶的差分来说,误差不断增长且具有二项式系数的形式.

3.11 计算表 3.4 中前二列所展示的函数的差分.这可视为一个误差函数类型,它的每一个值均是等于一个单位的舍入误差.试证:交错正负型导致在较高阶差分中严重的误差增长.幸好,舍入误差很少恰以如此方式交替变化.

解 ☞ 表 3.4 的其他列中出现的是一些所求的差分,对于每一较高阶的差分来说,其误差是低一阶差分的两倍.

表 3.4

x_0	1						
		-2					
x_1	-1		4				
		2		-8			
x_2	1		-4		16		
		-2		8		-32	
x_3	-1		4		-16		64
		2		-8		32	
x_4	1		-4		16		
		-2		8			
x_5	-1		4				
		2					
x_6	1						

3.12 一览表

$$1 \quad 2 \quad 4 \quad 8 \quad 16 \quad 26 \quad 42 \quad 64 \quad 93$$

中有一个数字被印错,是哪一个?

解 为了修改一览表,计算第一个4阶差分,并水平地展开它们,我们有

1	2	4	8	10	16	22	29
1	2	4	2	6	6	7	
1	2	2	4	0	1		
1	4	6	−4	1			

于是,这种想法是不可避免的:这些二项式的系数产生于原始表的中间项16中大小为1的数据误差,将它改为15产生新的表

1　2　4　8　15　26　42　64　93

由此,我们得到差分

1	2	4	7	11	16	22	29
1	2	3	4	5	6	7	

这意味着事情做得不坏.这是数据修均(date smoothing)的一个非常简单的例子.在后面的章节中,我们将会更充分地讨论它.始终存在着这种可能性:正如我们的原始表那样,数据来自一个不平整而非光滑的过程,以至于突起处(16而不是15)是真实数据而非印刷错误.上述分析可视为检测出了突起处,而非更正印刷错误.

补　充　题

3.13 对于如下 y_k 值,计算到4阶差分(这里可假设 $x_k = k$):

k	0	1	2	3	4	5	6
y_k	0	1	16	81	256	625	1296

3.14 对于 $k=5$,由定义直接证明

$$\Delta^5 y_0 = y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0$$

来验证题3.5.

3.15 仿照题3.9,证明 $\Delta \dfrac{u_k}{v_k} = \dfrac{v_k \Delta u_k - u_k \Delta v_k}{v_{k+1} v_k}$.

3.16 计算到5阶差分,以便观察大小为1的"相邻误差"的影响.

k	0	1	2	3	4	5	6	7
y_k	0	0	0	1	1	0	0	0

3.17 在如下 y_k 值中,找出一个单一误差并更正之.

k	0	1	2	3	4	5	6	7
y_k	0	0	1	6	24	60	120	210

3.18 利用线性性质证明:若 $y_k = k^3$,则

$$\Delta y_k = y_{k+1} - y_k = 3k^2 + 3k + 1,$$
$$\Delta^2 y_k = \Delta y_{k+1} - \Delta y_k = 6k + 6,$$
$$\Delta^3 y_k = \Delta^2 y_{k+1} - \Delta^2 y_k = 6.$$

3.19 试证:若 $y_k = k^4$,则 $\Delta^4 y_k = 24$.

3.20 试证：若 $y_k = 2^k$，则 $\Delta y_k = y_k$.

3.21 试证：若 $y_k = C^k$，则 $\Delta y_k = C^k(C-1)$.

3.22 由以下提供的一阶差分，算出缺省的 y_k 值.

y_k	0		·		·		·		·		·		·
Δy_k		1		2		4		7		11		16	

3.23 由以下提供的数据，算出缺省的 y_k 值与 Δy_k 值.

y_k	·		·		·		6		·		·		·
Δy_k		·		·		5		·		·		·	
$\Delta^2 y_k$			1		4		13		18		24		

3.24 由以下提供的数据，算出缺省的 y_k 值.

y_k	0		0		0		6		24		60		·		·		·
Δy_k		0		0		6		18		36		·		·		·	
$\Delta^2 y_k$			0		6		12		18		·		·		·		
$\Delta^3 y_k$				6		6		6		6		6		6			

3.25 在数据

y_k	1	3	11	31	69	113	223	351	521	739	1011

中找出一个印刷错误并更正之.

3.26 提升公式 $\Delta^2 y_0 = y_2 - 2y_1 + y_0$ 中所有的下标，对 $\Delta^2 y_1$ 和 $\Delta^2 y_2$ 写出类似的展开式，计算这些二阶差分的和，它应该等于 $\Delta y_3 - \Delta y_0 = y_4 - y_3 - y_1 + y_0$.

3.27 找出一个满足 $\Delta y_k = 2y_k$ 的函数 y_k.

3.28 找出一个满足 $\Delta^2 y_k = 9y_k$ 的函数 y_k. 你能找出两个这样的函数吗？

3.29 继续前题：寻找一个函数使得 $\Delta^2 y_k = 9k_2$ 并有 $y_0 = 0, y_1 = 1$

3.30 证明 $\Delta(\sin k) = 2\sin\frac{1}{2}\cos(k+1)$.

3.31 证明 $\Delta(\cos k) = -2\sin\frac{1}{2}\sin\left(k+\frac{1}{2}\right)$.

3.32 证明 $\Delta(\log x_k) = \log(1 + k/x_k)$，其中 $x_k = x_0 + kh$.

第四章　阶乘多项式

阶乘多项式

阶乘多项式(factorial polynomial)由

$$y_k = k^{(n)} = k(k-1)\cdots(k-n+1)$$

所定义,其中 n 是正整数.例如,$k^{(2)}=k(k-1)=k^2-k$.由于这些多项式具有使用方便的特性,故它们在有限差分的理论中起着重要作用.一个阶乘多项式的各种差分仍是阶乘多项式,更详细地说,一阶差分

$$\Delta k^{(n)} = nk^{(n-1)}$$

让人联想起微分是怎样作用于 x 的幂的.较高阶的差分则变成进一步降阶的阶乘多项式,直至最后

$$\Delta^n k^{(n)} = n!,$$

以及更高阶的差分全为零.

二项式系数以

$$\binom{k}{n} = \frac{k^{(n)}}{n!}$$

与阶乘多项式相关联.因此就分享这些多项式的某些性质.特别著名的递推式

$$\binom{k+1}{n+1} - \binom{k}{n+1} = \binom{k}{n}$$

具有一个有限差分公式的形式.

从阶乘多项式的定义可直接得到简单的递归式

$$k^{(n+1)} = (k-n)k^{(n)},$$

将它改写为

$$k^{(n)} = \frac{k^{(n+1)}}{k-n},$$

它能用于将阶乘思想推广到整数 $n=0,-1,-2,\cdots$.于是,基本公式

$$\Delta k^{(n)} = nk^{(n-1)}$$

对所有整数成立.

Stirling 数

当阶乘多项式用标准多项式的形式表示时,产生**第一类 Stirling 数**.于是

$$k^{(n)} = S_1^{(n)}k + \cdots + S_n^{(n)}k^n = \Sigma S_i^{(n)}k^i$$

中,$S_i^{(n)}$ 是 Stirling 数.例如

$$k^{(3)} = 2k - 3k^2 + k^3,$$

这使得 $S_1^{(3)}=2$, $S_2^{(3)}=-3$, $S_3^{(3)}=1$.逆推公式

$$S_i^{(n+1)} = S_{i-1}^{(n)} - nS_i^{(n)}$$

能迅速构造出 Stirling 数的表格.

当 k 的幂被表示为阶乘多项式的组合时,产生**第二类 Stirling 数**.于是

$$k^n = s_1^{(n)}k^{(1)} + \cdots + s_n^{(n)}k^{(n)} = \sum s_i^{(n)}k^{(i)}$$

中,$s_i^{(n)}$ 是 Stirling 数.例如

$$k^3 = k^{(1)} + 3k^{(2)} + k^{(3)},$$

从而 $s_1^{(3)}=1, s_2^{(3)}=3, s_3^{(3)}=1$. 逆推公式

$$s_i^{(n+1)} = s_{i-1}^{(n)} - is_i^{(n)}$$

能迅速构造出这些 Stirling 数的表格. 基本定理指出，每个 k 的幂，作为阶乘多项式的组合，仅有一个这样的表达式. 这确保了第二类 Stirling 数惟一.

任意多项式的表示法

任意多项式作为阶乘多项式的组合的表示法是下一个自然的步骤. k 的每一个幂被如此表示，然后将它们组合起来，由刚才介绍的基本定理，这种表示是惟一的. 例如

$$k^2+2k+1=[k^{(2)}+k^{(1)}]+2k^{(1)}+1=k^{(2)}+3k^{(1)}+1.$$

根据先将这种多项式表示为阶乘多项式的组合，以及随后对求差分的各阶乘项使用我们的公式，将便于找出任意多项式的差分. 至此已可理解本章的一个主要定理，该定理指出：一个 n 次多项式的差分是另一个 $n-1$ 次多项式，这使得 n 次多项式的 n 阶差分是一个常数，而更高阶的差分是零.

题　解

4.1　考虑满足 $y_k=k(k-1)(k-2)$ 的特殊函数，并证明 $\Delta y_k=3k(k-1)$.

证 ☞

$$\Delta y_k = y_{k+1}-y_k=(k+1)k(k-1)-k(k-1)(k-2)$$
$$=[(k+1)-(k-2)]k(k-1)=3k(k-1).$$

对于开始的几个整数 k 值，同样结果的表格形式给在表 4.1 中.

表 4.1

k	y_k	Δy_k
0	0	
		0
1	0	
		0
2	0	
		6
3	6	
		18
4	24	
		36
5	60	

4.2　将题 4.1 一般化. 考虑特殊函数

$$y_k=k(k-1)\cdots(k-n+1)=k^{(n)},$$

(注意：上标不是幂). 对于 $n>1$，证明

$$\Delta y_k = nk^{(n-1)}.$$

此结果使我们猛然联想到 n 次幂函数的有关导数定理.

证 ☞

$$\Delta y_k = y_{k+1}-y_k=[(k+1)\cdots(k-n+2)]-[k\cdots(k-n+1)]$$
$$=[(k+1)-(k-n+1)]k(k-1)\cdots(k-n+2)=nk^{(n-1)}$$

4.3　证明：若 $y_k=k^{(n)}$，则 $\Delta^2 y_k=n(n-1)k^{(n-2)}$.

证 ☞ 将题 4.2 应用于 Δy_k 而不是用于 y_k：

$$\Delta^2 k^{(n)}=\Delta\Delta k^{(n)}=\Delta nk^{(n-1)}=n(n-1)k^{(n-2)}.$$

推广至高阶差分正如求导那样进行.

4.4　证明 $\Delta^n k^{(n)}=n!$ 而 $\Delta^{n+1}k^{(n)}=0$.

证 ☞ 将题 4.2 应用 n 次以后立得第一个结果(符号 $k^{(0)}$可看作 1). 由于 $n!$ 是常数(与 k 无关), 它的差分全为零.

4.5 二项式系数是整数

$$\binom{k}{n} = \frac{k^{(n)}}{n!} = \frac{k!}{n!(k-n)!}.$$

证明递推公式

$$\binom{k+1}{n+1} = \binom{k}{n+1} + \binom{k}{n}.$$

证 ☞ 利用阶乘多项式以及应用题 4.2

$$\binom{k+1}{n+1} - \binom{k}{n+1} = \frac{(k+1)^{(n+1)}}{(n-1)!} - \frac{k^{(n+1)}}{(n+1)!} = \frac{\Delta k^{(n+1)}}{(n+1)!}$$

$$= \frac{(n+1)k^{(n)}}{(n+1)!} = \frac{k^{(n)}}{n!} = \binom{k}{n}.$$

移项后可立即证出结果. 这个著名的结果已被使用过.

4.6 利用递推公式, 将直到 $k=8$ 的二项式系数制成表格.

解 ☞ 表 4.2 的第一列给出的 $\binom{k}{0}$ 被定义为 1. 根据定义, 对角线 $k=n$ 处是 1, 其他表值来源于递推公式. 此表易于推广.

表 4.2

k \ n	0	1	2	3	4	5	6	7	8
1	1	1							
2	1	2	1						
3	1	3	3	1					
4	1	4	6	4	1				
5	1	5	10	10	5	1			
6	1	6	15	20	15	6	1		
7	1	7	21	35	35	21	7	1	
8	1	8	28	56	70	56	28	8	1

4.7 试证: 若 k 是一个正整数, 则当 $n>k$ 时, $k^{(n)}$ 与 $\binom{k}{n}$ 是零 $\left[\text{对于 } n>k, \text{符号} \binom{k}{n} \text{定义为} \dfrac{k^{(n)}}{n!}\right]$.

证 ☞ 注意到 $k^{(k+1)}=k(k-1)\cdots 0$, 当 $n>k$, 阶乘 $k^{(n)}$将包含零因子, 从而 $\binom{k}{n}$ 是零.

4.8 二项式系数符号和阶乘符号常用于非整数 k. 对于 $k=\frac{1}{2}$ 和 $n=2, 3$, 计算 $k^{(n)}$ 及 $\binom{k}{n}$.

解 ☞

$$k^{(2)} = \left(\frac{1}{2}\right)^{(2)} = \frac{1}{2}\left(\frac{1}{2}-1\right) = -\frac{1}{4},$$

$$k^{(3)} = \left(\frac{1}{2}\right)^{(3)} = \frac{1}{2}\left(\frac{1}{2}-1\right)\left(\frac{1}{2}-2\right) = \frac{3}{8},$$

$$\binom{k}{2} = \frac{k^{(2)}}{2!} = \frac{1}{2}\left(-\frac{1}{4}\right) = -\frac{1}{8},$$

$$\binom{k}{3} = \frac{k^{(3)}}{3!} = \frac{1}{6}\left(\frac{3}{8}\right) = \frac{1}{16}.$$

4.9 阶乘的思想也可推广到上标不是正整数的情况. 根据定义, 当 n 是一个正整数时, 有 $k^{(n+1)}=(k-n)k^{(n)}$. 将它改写为

$$k^{(n)} = \frac{1}{k-n}k^{(n+1)},$$

并用作 $n=0, -1, -2, \cdots$ 时 $k^{(n)}$的定义. 试证 $k^{(0)}=1$ 且 $k^{(-n)}=1/(k+n)^{(n)}$.

证 ☞ 将 $n=0$ 代入立即有第一个结果. 对于第二个结果,我们有

$$k^{(-1)}=\frac{1}{k+1}k^{(0)}=\frac{1}{k+1}=\frac{1}{(k+1)^{(1)}},$$

$$k^{(-2)}=\frac{1}{k+2}k^{(-1)}=\frac{1}{(k+2)(k+1)}=\frac{1}{(k+2)^{(2)}},$$

如此等等,这需要用归纳法来证明,而此处省略其细节. 对于 $k=0$, 定义 $k^{(0)}=1$ 并接受此结果往往是方便的.

4.10 证明:对于所有的整数 n, $\Delta k^{(n)}=nk^{(n-1)}$.

证 ☞ 对于 $n>1$, 已在题 4.2 中证明. 对于 $n=1$ 和 0, 可以立即得到结论. 对于 n 为负整数,比如 $n=-p$,

$$\Delta k^{(n)}=\Delta k^{(-p)}=\Delta\frac{1}{(k+p)^{(p)}}=\frac{1}{(k+1+p)\cdots(k+2)}-\frac{1}{(k+p)\cdots(k+1)}$$

$$=\frac{1}{(k+p)\cdots(k+2)}\left(\frac{1}{k+1+p}-\frac{1}{k+1}\right)=\frac{-p}{(k+1+p)\cdots(k+1)}$$

$$=\frac{n}{(k+1-n)^{(1-n)}}=nk^{(n-1)}.$$

这个结果类似于微积分学中"若 $f(x)=x^n$, 则对于所有的整数, $f'(x)=nx^{n-1}$ 均成立"这一定理.

4.11 求 $\Delta k^{(-1)}$.

解 ☞ 由前题知

$$\Delta k^{(-1)}=-k^{(-2)}=\frac{-1}{(k+2)(k+1)}.$$

4.12 试证 $k^{(2)}=-k+k^2$, $k^{(3)}=2k-3k^2+k^3$, $k^{(4)}=-6k+11k^2-6k^3+k^4$.

证 ☞ 直接从定义出发:

$$k^{(2)}=k(k-1)=-k+k^2,$$

$$k^{(3)}=k^{(2)}(k-2)=2k-3k^2+k^3,$$

$$k^{(4)}=k^{(3)}(k-3)=-6k+11k^2-6k^3+k^4.$$

4.13 将题 4.12 推广,试证:在一个阶乘多项式表示为标准多项式的展开式

$$k^{(n)}=S_1^{(n)}k+\cdots+S_n^{(n)}k^n=\sum S_i^{(n)}k^i$$

中,系数满足递推公式

$$S_i^{(n+1)}=S_{i-1}^{(n)}-nS_i^{(n)}.$$

这些系数被称为**第一类 Stirling 数**.

证 ☞ 以 $n+1$ 代替 n,

$$k^{(n+1)}=S_1^{(n+1)}k+\cdots+S_{n+1}^{(n+1)}k^{n+1},$$

而利用 $k^{(n+1)}=k^{(n)}(k-n)$ 这一事实,我们有

$$S_1^{(n+1)}k+\cdots+S_{n+1}^{(n+1)}k^{n+1}=[S_1^{(n)}k+\cdots+S_n^{(n)}k^n](k-n).$$

今比较式子两端的 k^i 的系数,有

$$S_i^{(n+1)}=S_{i-1}^{(n)}-nS_i^{(n)},\quad i=2,\cdots,n.$$

由比较 k 及 k^{n+1} 的系数,还应注意到特殊情况: $S_1^{(n+1)}=-nS_1^{(n)}$, $S_{n+1}^{(n+1)}=S_n^{(n)}$.

4.14 利用题 4.13 中的公式构造一个第一类 Stirling 数的简表.

解 ☞ 由特殊情况下的公式 $S_1^{(n+1)}=-nS_1^{(n)}$ 立即导出表 4.3 的第一列. 例如,因为 $S_1^{(1)}$ 显然是 1, 故

$$S_1^{(2)}=-S_1^{(1)}=-1,\ S_1^{(3)}=-2S_1^{(2)}=2,$$

如此等等. 另一特殊情况下的公式将表的顶部对角线全填上 1, 然后,我们的主要递推公式完成了此表. 例如

$$S_2^{(3)}=S_1^{(2)}-2S_2^{(2)}=(-1)-2(1)=-3,$$

$$S_2^{(4)} = S_1^{(3)} - 3S_2^{(3)} = (2) - 3(-3) = 11,$$
$$S_3^{(4)} = S_2^{(3)} - 3S_3^{(3)} = (-3) - 3(1) = -6,$$

如此等等.直至 $n=8$,该表读取如下:

表 4.3

n \ i	1	2	3	4	5	6	7	8
1	1							
2	-1	1						
3	2	-3	1					
4	-6	11	6	1				
5	24	50	35	-10	1			
6	-120	274	225	85	-15	1		
7	720	-1 764	1 624	-735	175	-21	1	
8	-5 040	13 068	-13 132	6 769	-1 960	322	-28	1

4.15 利用表 4.3 展开 $k^{(5)}$.

解 ☞ 利用表的第 5 行,

$$k^{(5)} = 24k - 50k^2 + 35k^3 - 10k^4 + k^5.$$

4.16 试证

$$k^2 = k^{(1)} + k^{(2)}, \quad k^3 = k^{(1)} + 3k^{(2)} + k^{(3)}, k^4 = k^{(1)} + 7k^{(2)} + 6k^{(3)} + k^{(4)}.$$

证 ☞ 利用表 4.3,

$$k^{(1)} + k^{(2)} = k + (-k + k^2) = k^2,$$
$$k^{(1)} + 3k^{(2)} + k^{(3)} = k + 3(-k + k^2) + (2k - 3k^2 + k^3) = k^3,$$
$$k^{(1)} + 7k^{(2)} + 6k^{(3)} + k^{(4)} = k + 7(-k + k^2) + 6(2k - 3k^2 + k^3)$$
$$+ (-6k + 11k^2 - 6k^3 + k^4) = k^4.$$

4.17 作为下题的必要准备,证明 k 的幂作为阶乘多项式的组合仅有一个表示法.

证 ☞ 对于 k^p,假设存在两个这样的表示法,

$$k^p = A_1k^{(1)} + \cdots + A_pk^{(p)}, \quad k^p = B_1k^{(1)} + \cdots + B_pk^{(p)}.$$

两式相减有

$$0 = (A_1 - B_1)k^{(1)} + \cdots + (A_p - B_p)k^{(p)}.$$

由于式子的右端是一个多项式,而没有多项式可对所有的 k 值为零,于是,右端 k 的每一幂的系数必须是零.但 k^p 仅会出现在最后一项中,因此一定有 $A_p = B_p$.然后,k^{p-1}仅会出现在剩余项的最后一项中,这将是$(A_{p-1} - B_{p-1})k^{(p-1)}$项,因此有 $A_{p-1} = B_{p-1}$.同理,直至 $A_1 = B_1$.

这个证明是对惟一表示法的典型证明方法,它在数值分析中被频繁地采用.类似的定理:两个多项式没有恒等的系数就不可能有恒等的值,是代数中的经典结论,已被用于题 4.13.

4.18 将题 4.16 推广到一般,试证 k 的幂能表示为阶乘多项式的组合

$$k^n = s_1^{(n)}k^{(1)} + \cdots + s_n^{(n)}k^{(n)} = \sum s_i^{(n)}k^{(i)},$$

并且系数满足递推公式

$$s_i^{(n+1)} = s_{i-1}^{(n)} + is_i^{(n)}.$$

这些系数被称为**第二类 Stirling 数**.

解 ☞ 我们继续利用归纳法.对于小的 k,题 4.16 已确立了这种表示法的存在性.假设

$$k^n = s_1^{(n)}k^{(1)} + \cdots + s_n^{(n)}k^{(n)},$$

乘以 k 以后得到

$$k^{n+1} = ks_1^{(n)}k^{(1)} + \cdots + ks_n^{(n)}k^{(n)}.$$

今注意到

$$k \cdot k^{(i)} = (k-i)k^{(i)} + ik^{(i)} = k^{(i+1)} + ik^{(i)}$$

从而

$$k^{n+1} = s_1^{(n)}(k^{(2)} + k^{(1)}) + \cdots + s_n^{(n)}(k^{(n+1)} + nk^{(n)})$$

这已是k^{n+1}的一个表示法,从而,我们能写出

$$k^{n+1} = s_1^{(n+1)}k^{(1)} + \cdots + s_{n+1}^{(n+1)}k^{(n+1)}.$$

以完成归纳法.根据题4.17,在最后排齐了的两式中,$k^{(i)}$的系数必定相同,从而

$$s_i^{(n+1)} = s_{i-1}^{(n)} + is_i^{(n)}, \quad i = 2, \cdots, n.$$

比较$k^{(1)}$和$k^{(n+1)}$的系数还可注意到如下特殊情况:

$$s_1^{(n+1)} = s_1^{(n)}, \quad s_{n+1}^{(n+1)} = s_n^{(n)}.$$

4.19 利用题4.18中的公式构造一个第二类Stirling数的简表.

解 ☞ 由于$s_1^{(1)}$显然是1,由特殊公式$s_1^{(n+1)} = s_1^{(n)}$立即导出表4.4的第一列.由另一特殊公式得到顶部对角线,然后,用主要的递推公式完成此表.例如

$$s_2^{(3)} = s_1^{(2)} + 2s_2^{(2)} = (1) + 2(1) = 3,$$

$$s_2^{(4)} = s_1^{(3)} + 2s_2^{(3)} = (1) + 2(3) = 7,$$

$$s_3^{(4)} = s_2^{(3)} + 3s_3^{(3)} = (3) + 3(1) = 6,$$

等等,直至$n=8$.该表读取如下:

表 4.4

n \ i	1	2	3	4	5	6	7	8
1	1							
2	1	1						
3	1	3	1					
4	1	7	6	1				
5	1	15	25	10	1			
6	1	31	90	65	15	1		
7	1	63	301	350	140	21	1	
8	1	127	966	1701	1050	266	28	1

4.20 利用表4.4,用阶乘多项式展开k^5.

解 ☞ 利用该表的第5行,

$$k^5 = k^{(1)} + 15k^{(2)} + 25k^{(3)} + 10k^{(4)} + k^{(5)}.$$

4.21 证明:一个n次多项式的第n阶差分是相等的,高于第n阶的差分是零.

证 ☞ 记此多项式为$P(x)$,考虑它在等距自变量$x_0, x_1, x_2, \cdots$这个离散集上的值.我们经常使用的以整数变量k来代替x的处理方法常常是很方便的,这个k由$x_k - x_0 = kh$与x产生联系,其中,h是连续变量x之间的等差.对于自变量k,用符号P_k表示我们的关于自变量k的多项式的值.由于这种自变量的代换是线性的,故以x与以k来表示,多项式均具有相同的次,我们可将它写为

$$P_k = a_0 + a_1k + a_2k^2 + \cdots + a_nk^n.$$

题4.18已证明了k的每一个幂能被表示为阶乘多项式的一个组合,这导致P_k自身作为这种组合的一个表示:

$$P_k = b_0 + b_1k^{(1)} + b_2k^{(2)} + \cdots + b_nk^{(n)}.$$

利用题4.2及线性性质,

$$\Delta P_k = b_1 + 2b_2k^{(1)} + \cdots + nb_nk^{(n-1)},$$

反复使用题4.2,最后导出$\Delta^n P_k = n!\,b_n$.于是,所有的第n阶差分都是这个数,它们不随k变化,而更

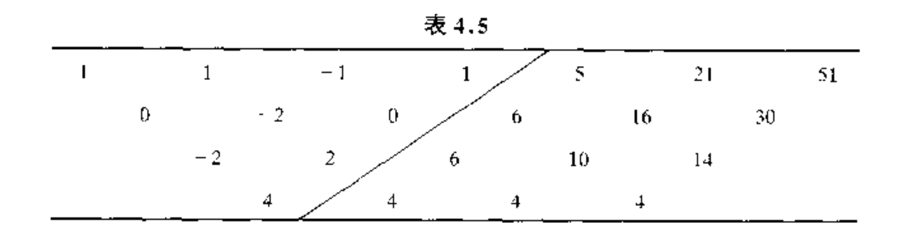
高阶的差分是零.

4.22 假设如下的 y_k 值属于一个 4 次多项式,计算其余的 3 个值.

k	0	1	2	3	4	5	6	7
y_k	0	1	2	1	0	·	·	·

解 ☞ 根据题 4.21,一个 4 次多项式第 4 阶差分是常数.从给定的数据出发进行计算,我们得到表 4.5 中斜线左边的表值.

表 4.5



1		1		−1		1		5		21		51
	0		−2		0		6		16		30	
		−2		2		6		10		14		
			4		4		4		4			

假设其余的 4 阶差分也是 4,则导出线右边的表值,由此可料想缺少的表值为 $y_5=5$, $y_6=26$, $y_7=77$.

补 充 题

4.23 计算阶乘:$6^{(3)}, 6^{(6)}, 6^{(7)}, \left(\frac{1}{3}\right)^{(2)}, \left(\frac{1}{3}\right)^{(3)}, \left(\frac{1}{3}\right)^{(4)}$.

4.24 计算阶乘:$6^{(-1)}, 6^{(-2)}, 6^{(-3)}, \left(\frac{1}{3}\right)^{(-1)}, \left(\frac{1}{3}\right)^{(-2)}, \left(\frac{1}{3}\right)^{(-3)}$.

4.25 计算二项式系数:$\binom{6}{3}, \binom{6}{6}, \binom{6}{7}, \binom{\frac{1}{3}}{2}, \binom{\frac{1}{3}}{3}, \binom{\frac{1}{3}}{4}$.

4.26 对于 $y_k=k^{(4)}$ 的这些值,计算到 4 阶差分.

k	0	1	2	3	4	5	6	7
y_k	0	0	0	0	24	120	360	840

4.27 借助于阶乘多项式,利用题 4.2 表示 $y_k=k^{(4)}$ 的开头 4 个差分.

4.28 借助于阶乘多项式,利用题 4.2 表示 $y_k=k^{(5)}$ 的开头 5 个差分.

4.29 利用表 4.3 将 $y_k=2k^{(3)}-k^{(2)}+4k^{(1)}-7$ 表示为一个普通多项式.

4.30 利用表 4.3 将 $y_k=k^{(6)}+k^{(3)}+1$ 表示为一个普通多项式.

4.31 利用表 4.4 将 $y_k=\frac{1}{3}(2k^4-8k^2+3)$ 表示为阶乘多项式的一个组合.

4.32 利用表 4.4 将 $y_k=80k^3-30k^4+3k^5$ 表示为阶乘多项式的一个组合.

4.33 借助于阶乘多项式,利用上述问题的结果算出 Δy_k,然后,利用表 4.3 将该结果转换到普通多项式.

4.34 借助于阶乘多项式,利用题 4.32 的结果得出 Δy_k 和 $\Delta^2 y_k$,然后利用表 4.3 将两结果转换为普通多项式.

4.35 假设如下的 y_k 是一个 4 次多项式的值,推测接下去的 3 个值.

k	0	1	2	3	4	5	6	7
y_k	1	−1	1	1	1			

4.36　假设如下的 y_k 是一个 4 次多项式的值，推测接下去的 3 个值.

k	0	1	2	3	4	5	6	7
y_k	0	0	1	0	0			

4.37　对于取如下值的一个多项式来说，可能的最低次是多少？

k	0	1	2	3	4	5
y_k	0	3	8	15	24	35

4.38　对于取如下值的一个多项式来说，可能的最低次是多少？

k	0	1	2	3	4	5
y_k	0	1	1	1	1	0

4.39　找出一个函数 y_k，对它而言 $\Delta y_k = k^{(2)} = k(k-1)$.

4.40　找出一个函数 y_k，对它而言 $\Delta y_k = k(k-1)(k-2)$.

4.41　找出一个函数 y_k，对它而言 $\Delta y_k = k^2 = k^{(2)} + k^{(1)}$.

4.42　找出一个函数 y_k，对它而言 $\Delta y_k = k^3$.

4.43　找出一个函数 y_k，对它而言 $\Delta y_k = 1(k+1)(k+2)$.

第五章 求 和 法

正如积分法相对于微分法,求和法(summation)是相对于差分化(differencing)的逆运算.广泛的讨论将出现在第 17 章,而此处则介绍两个基本的结果.

1. **嵌套和**(telescoping sum)是差分的和,我们具有简单而有用的

$$\sum_{k=0}^{n-1}\Delta y_k = y_n - y_0.$$

这类似于导数的积分法.倘若对于函数 y_k,方程 $\Delta y_k = z_k$ 能求解出来,那么任意和能转换为嵌套和.于是

$$\sum_{k=0}^{n-1} z_k = \sum_{k=0}^{n-1}\Delta y_k = y_n - y_0.$$

有限积分法(finite integration)是从

$$\Delta y_k = z_k$$

获得 y_k 的过程,其中 z_k 是已知的.因为显然有

$$y_n = y_0 + \sum_{k=0}^{n-1} z_k,$$

故有限积分法与求和法是相同的问题.然而,正如在积分计算中那样,存在许多显式有限积分(不含 Σ)有用的机会.

2. **分部求和法**(summatation by parts)是求和法运算的另一主要结果,它包括公式

$$\sum_{i=0}^{n-1} u\Delta v_i = u_n v_n - u_0 v_0 - \sum_{i=0}^{n-1} v\Delta u_i,$$

这类似于相应的分部积分公式.

这个公式的应用包含着以一个(推测上)比较简单的和去取代原先的和.若两个 Σ 中有一个是已知的,则该公式起着由一个和定出另一个和的作用.当嵌套和或分部求和法能适合于无穷级数的部分和时,该无穷级数也能被求值.

题 解

5.1 证明 $\sum\limits_{k=0}^{n-1}\Delta y_k = y_n - y_0$.

证 ☞ 这是一个简单而有用的结果.由于它包括差分的求和法,所以通常将它与包括一个导数的积分法这一类似于微积分的结果相比较.首先注意到

$$\Delta y_0 = y_1 - y_0,$$

$$\Delta y_0 + \Delta y_1 = (y_1 - y_0) + (y_2 - y_1) = y_2 - y_0,$$

$$\Delta y_0 + \Delta y_1 + \Delta y_2 = (y_1 - y_0) + (y_2 - y_1) + (y_3 - y_2) = y_3 - y_0,$$

这些式子描述了所含的这种嵌套和.一般地,

$$\sum_{k=0}^{n-1}\Delta y_k = (y_1 - y_0) + (y_2 - y_1) + (y_3 - y_2) + \cdots + (y_n - y_{n-1}) = y_n - y_0$$

所有其他的 y 值既出现正号,又出现负号.在差分表中观察,结果看上去甚至更简单.邻近的差分之和给出了上行中两个表值的差

$$\begin{matrix} y_0 & & \cdot & & \cdot & & \cdot & & \cdot & & \cdot & & \cdot & & y_n \\ & \Delta y_0 & & \Delta y_1 & & \Delta y_2 & & \cdot & & \cdot & & \cdot & & \Delta y_{n-1} & \end{matrix}$$

在表的其他地方仍有类似结果.

5.2 证明 $1^2+2^2+\cdots+n^2=\sum_{i=1}^{n} i^2=\frac{n(n+1)(2n+1)}{6}$.

证 ☞ 我们需要求一个函数，对它而言 $\Delta y_i=i^2$；这类似于微积分中的积分问题. 在这个简单的例子中，y_i 几乎能凭直觉找出，尽管如此，我们仍使用一个也能处理难题的方法. 首先，以一个阶乘多项式的组合代换 i^2；利用 Stirling 数，

$$\Delta y_i = i^2 = i^{(2)} + i^{(1)},$$

具有这个差分的一个函数是

$$y_i = \frac{1}{3}i^{(3)} + \frac{1}{2}i^{(2)},$$

这容易由计算 Δy_i 来证实. 由 Δy_i 得到 y_i 被称为**有限积分法**. 显然，它类似于导数的积分. 现改写题5.1的结果为 $\sum_{i=1}^{n}\Delta y_i = y_{n+1}-y_1$. 经过代换得到

$$\sum_{i=1}^{n} i^2 = \left[\frac{1}{3}(n+1)^{(3)}+\frac{1}{2}(n+1)^{(2)}\right]-\left[\frac{1}{3}(1)^{(3)}+\frac{1}{2}(1)^{(2)}\right]$$

$$=\frac{(n+1)n(n-1)}{3}+\frac{(n+1)n}{2}=\frac{n(n+1)(2n+1)}{6}.$$

5.3 求级数 $\sum_{i=0}^{\infty}\frac{1}{(i+1)(i+2)}$ 的值.

解 ☞ 由以前的结果 $\Delta i^{(-1)}=\frac{-1}{(i+1)(i+2)}$，利用题4.9处理 $0^{(-1)}$，则

$$S_n=\sum_{i=0}^{n-1}\frac{1}{(i+1)(i+2)}=-\sum_{i=0}^{n-1}\Delta i^{(-1)}=-\left[n^{(-1)}-0^{(-1)}\right]$$

$$=1-\frac{1}{n+1}.$$

原级数定义为 $\lim S_n$，它因此等于1.

5.4 考虑定义在同一个自变量 x_k 集合上的两个函数，它们分别有值 u_k 与 v_k. 证明

$$\sum_{i=0}^{n-1}u_i\Delta v_i = u_nv_n-u_0v_0-\sum_{i=0}^{n-1}v_{i+1}\Delta u_i.$$

证 ☞ 这被称为**分部求和法**，它类似于微积分中的结果

$$\int_{x_0}^{x_n}u(x)v'(x)dx = u(x_n)v(x_n)-u(x_0)v(x_0)-\int_{x_0}^{x_n}v(x)u'(x)dx.$$

证明始于题3.9的结果；稍作整理有

$$u_i\Delta v_i = \Delta(u_iv_i) - v_{i+1}\Delta u_i,$$

从 $i=0$ 加到 $i=n-1$，有

$$\sum_{i=0}^{n-1}u_i\Delta v_i = \sum_{i=0}^{n-1}\Delta(u_iv_i)-\sum_{i=0}^{n-1}v_{i+1}\Delta u_i,$$

然后对右端的第一个和式应用题5.1. 接着就得到所需的结果.

5.5 求级数 $\sum_{i=0}^{\infty}iR^i$ 的值，其中 $-1<R<1$.

解 ☞ 由于 $\Delta R^i=R^{i+1}-R^i=R^i(R-1)$，我们可设 $u_i=i$，$v_i=R^i/(R-1)$，并利用分部求和法. 取有限和

$$S_n=\sum_{i=0}^{n-1}iR^i=\sum_{i=0}^{n-1}u_i\Delta v_i = n\cdot\frac{R^n}{R-1}-0-\sum_{i=0}^{n-1}\frac{R^{i+1}}{R-1},$$

最后的和式是一个几何级数且符合一个基本公式，从而

$$S_n=\frac{nR^n}{R-1}+\frac{R(1-R^n)}{(1-R)^2}.$$

由于 nR^n 与 R^{n+1} 极限均为零，故该无穷级数的值是

$$\lim S_n = R/(1-R)^2.$$

5.6 投掷一枚硬币，直至第一次显示正面朝上，若在第 i 次投掷时第一次显示正面朝上，则形成一个等于 i 美元的支付(payoff)(若在第一次投掷时就立即显示正面朝上则支付 1 美元，在第二次投掷时显示正面朝上就支付 2 美元，依此类推). 对于平均支付(average payoff)，概率论导出如下级数：

$$1\left(\frac{1}{2}\right)+2\left(\frac{1}{4}\right)+3\left(\frac{1}{8}\right)+\cdots=\sum_{i=0}^{\infty} i\left(\frac{1}{2}\right)^{i}.$$

试利用上题中的结果计算此级数.

解 ☞ 由题 5.5，取 $R=\frac{1}{2}$，得 $\sum_{i=0}^{\infty} i\left(\frac{1}{2}\right)^{i}=\left(\frac{1}{2}\right)\Big/\left(\frac{1}{4}\right)=2$ 美元.

5.7 将分部求和法用于求级数 $\sum_{i=0}^{\infty} i^2R^i$ 的值.

解 ☞ 设 $u_i=i^2, v_i=R^i/(R-1)$，我们得到 $\Delta u_i=2i+1$，从而

$$S_n=\sum_{0}^{n-1} i^2R^i=\sum_{i=0}^{n-1} u_i\Delta v_i=n^2\frac{R^n}{R-1}-0-\sum_{i=0}^{n-1}\frac{R^{i+1}}{R-1}(2i+1)$$

$$=n^2\frac{R^n}{R-1}-\frac{2R}{R-1}\sum_{i=0}^{n-1} iR^i-\frac{R}{R-1}\sum_{i=0}^{n-1}R^i.$$

在题 5.5 中已求出了剩余的两个和式中的第一个，而第二个和式是几何级数，于是我们得到

$$S_n=\frac{n^2R^n}{R-1}-\frac{2R}{R-1}\left[\frac{nR^n}{R-1}+\frac{R(1-R^n)}{(1-R)^2}\right]-\frac{R}{R-1}\cdot\frac{1-R^n}{1-R}.$$

令 $n\to\infty$，最后得到 $\lim S_n=(R+R^2)/(1-R)^3$.

5.8 投掷一枚硬币，直至第一次显示正面朝上. 若在第 i 次投掷时第一次显示正面朝上，则形成一个等于 i^2 美元的支付. 对于平均支付，概率论导出级数 $\sum_{i=0}^{\infty} i^2\left(\frac{1}{2}\right)^{i}$. 求此级数的值.

解 ☞ 由题 5.7，取 $R=\frac{1}{2}$，得 $\sum_{i=0}^{\infty} i^2\left(\frac{1}{2}\right)^{i}=\left(\frac{1}{2}+\frac{1}{4}\right)\Big/\left(\frac{1}{8}\right)=6$ 美元.

补 充 题

5.9 利用有限积分法(如同在题 5.2 中那样)证明

$$\sum_{i=1}^{n} i=1+2+\cdots+n=\frac{n(n+1)}{2}.$$

5.10 由有限积分法求 $\sum_{i=1}^{n} i^3$ 的值.

5.11 利用有限积分法，试证 $\sum_{i=0}^{n-1} A^i=\frac{A^n-1}{A-1}$(参看题 3.21) 显然，它就是初等代数中的几何和(geometric sum).

5.12 试证 $\sum_{i=1}^{n-1}\binom{i}{k}=\binom{n}{k+1}-\binom{1}{k+1}$.

5.13 用有限积分法求 $\sum_{i=0}^{\infty}\frac{1}{(i+1)(i+2)(i+3)}$ 的值.

5.14 求 $\sum_{i=1}^{\infty}\frac{1}{i(i+2)}$ 的值.

5.15 对于 $-1<R<1$，求 $\sum_{i=0}^{\infty} i^3R^i$ 的值.

5.16 变更题 5.8，使得支付为 i^3. 利用题 5.15 求平均支付 $\sum_{i=0}^{\infty} i^3\left(\frac{1}{2}\right)^{i}$ 的值.

5.17 变更题 5.8，使得当 i 为偶数时支付是 $+1$，当 i 为奇数时支付为 -1. 平均支付是 $\sum_{i=1}^{\infty}(-1)^i\left(\frac{1}{2}\right)^{i}$，求该级数的值.

5.18 求 $\sum_{i=1}^{n} \lg\left(1 + \frac{1}{i}\right)$ 的值.

5.19 借助于 Stirling 数求 $\sum_{i=1}^{n} i^n$ 的值.

5.20 求 $\sum_{i=1}^{\infty} \left[\frac{1}{i(i+n)}\right]$ 的值.

5.21 求 $\sum_{i=0}^{\infty} i^n R^i$ 的值.

5.22 以一个求和的形式表示 $\Delta y_k = \frac{1}{k}$ 的一个有限积分,避免 $k = 0$.

5.23 以一个求和的形式表示 $\Delta y_k = \lg k$ 的有限积分.

第六章 Newton 公式

现可借助于有限差分与阶乘多项式表示配置多项式.首先证明求和公式

$$y_k = \sum_{i=0}^{k} \binom{k}{i} \Delta^i y_0,$$

并对配置多项式直接导出 **Newton 公式**.它可写为

$$p_k = \sum_{i=0}^{n} \binom{k}{i} \Delta^i y_0.$$

利用自变量 x_k,其中 $x_k = x_0 + kh$,能得到 Newton 公式的另一形式,并可证明它是

$$p(x_k) = y_0 + \left(\frac{\Delta y_0}{h}\right)(x_k - x_0) + \left(\frac{\Delta^2 y_0}{2!\,h^2}\right)(x_k - x_0)(x_k - x_1)$$

$$+ \cdots + \left(\frac{\Delta^n y_0}{n!\,h^n}\right)(x_k - x_0)\cdots(x_k - x_{n-1}),$$

其配置点是 $x_0, x_1, \cdots, x_n$.在自变量的这些点上,该多项式取指定的值 $y_0, y_1, \cdots, y_n$.

题 解

6.1 证明

$$y_1 = y_0 + \Delta y_0,$$
$$y_2 = y_0 + 2\Delta y_0 + \Delta^2 y_0,$$
$$y_3 = y_0 + 3\Delta y_0 + 3\Delta^2 y_0 + \Delta^3 y_0,$$

并推出如同

$$\Delta y_2 = \Delta y_0 + 2\Delta^2 y_0 + \Delta^3 y_0,$$
$$\Delta^2 y_2 = \Delta^2 y_0 + 2\Delta^3 y_0 + \Delta^4 y_0$$

这种类似的结果.

证 ☞ 就最一般结果而言,这仅仅是一个预备工作.第一个结果是显然的.对于第二个结果,利用表 6.1 一眼就可看出

$$y_2 = y_1 + \Delta y_1 = (y_0 + \Delta y_0) - (\Delta y_0 + \Delta^2 y_0),$$

表 6.1

x_0	y_0				
		Δy_0			
x_1	y_1		$\Delta^2 y_0$		
		Δy_1		$\Delta^3 y_0$	
x_2	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$
		Δy_2		$\Delta^3 y_1$	
x_3	y_3		$\Delta^2 y_2$		
		Δy_3			
x_4	y_4				

从而立即导出所需结果.注意到由表 6.1 最上面的对角线上的表值表示 y_2,同时注意到,几乎完全一样的计算得到

$$\Delta y_2 = \Delta y_0 + 2\Delta^2 y_0 + \Delta^3 y_0, \quad \Delta^2 y_2 = \Delta^2 y_0 + 2\Delta^3 y_0 + \Delta^4 y_0,$$

等等,由那些最上面的对角线上的表值表示 y_2 所在的对角线上的表值,最后,

$$y_3 = y_2 + \Delta y_2 = (y_0 + 2\Delta y_0 + \Delta^2 y_0) + (\Delta y_0 + 2\Delta^2 y_0 + \Delta^3 y_0)$$

立即导出第三个所需的结果.对于 $\Delta y_3, \Delta^2 y_3$ 等等,类似的表示能由仅仅提升每个 Δ 的上标而写出来.

6.2 证明:对于任一正整数 k, $y_k = \sum_{i=0}^{k} \binom{k}{i} \Delta^i y_0$ (这里,$\Delta^0 y_0$ 仅表示 y_0).

证 ☞ 利用归纳法来证明.对于 $k=1,2,3$,见题 6.1.假设当 k 是某个特别的整数 p 时,结果

$$y_p = \sum_{i=0}^{p} \binom{p}{i} \Delta^i y_0$$

成立,则正如前题中所指出的,我们的各种差分定义使得

$$\Delta y_p = \sum_{i=0}^{p} \binom{p}{i} \Delta^{i+1} y_0$$

也成立.此时我们得到

$$\begin{aligned} y_{p+1} = y_p + \Delta y_p &= \sum_{j=0}^{p} \binom{p}{j} \Delta^j y_0 + \sum_{j=1}^{p+1} \binom{p}{j-1} \Delta^j p_0 \\ &= y_0 + \sum_{j=1}^{p} \left[\binom{p}{j} + \binom{p}{j-1} \right] \Delta^j y_0 + \Delta^{p+1} y_0 \\ &= y_0 + \sum_{j=1}^{p} \binom{p+1}{j} \Delta^j y_0 + \Delta^{p+1} y_0 = \sum_{j=0}^{p+1} \binom{p+1}{j} \Delta^j y_0. \end{aligned}$$

第二步用到了题 4.5.如果有必要,和标可立即由 j 改为 i.从而,我们的结果当 $k = p+1$ 时成立.归纳法证毕.

6.3 证明:对于 $k = 0, 1, \cdots, n$, n 次多项式

$$\begin{aligned} p_k &= y_0 + k\Delta y_0 + \frac{1}{2!} k^{(2)} \Delta^2 y_0 + \cdots + \frac{1}{n!} k^{(n)} \Delta^n y_0 \\ &= \sum_{i=0}^{n} \frac{1}{i!} k^{(i)} \Delta^i y_0 = \sum_{i=0}^{n} \binom{k}{i} \Delta^i y_0 \end{aligned}$$

取值 $p_k = y_k$.这是 Newton 公式.

证 ☞ 首先,注意到当 $k=0$ 时,右端仅剩 y_0 项,而其余所有的项为零;当 $k=1$ 时,右端仅剩前两项,其余的为零;当 $k=2$ 时仅剩前三项.于是,利用题 6.1,

$$p_0 = y_0, \quad p_1 = y_0 + \Delta y_0 = y_1, \quad p_2 = y_0 + 2\Delta y_0 + \Delta^2 y_0 = y_2.$$

我们的证明本质已明.一般地,若 k 是从 0 至 n 的任一整数,则对于 $i > k$, $k^{(i)} = 0$(因为它将包含因子 $k-k$).该和缩写为

$$p_k = \sum_{i=0}^{k} \frac{1}{i!} k^{(i)} \Delta^i y_0,$$

而据题 6.2,它简化成 y_k.因此,这个问题的多项式对整数自变量 $k = 0, 1, \cdots, n$,与 y_k 函数取相同的值(不过,该多项式对任意自变量 k 都有定义).

6.4 以自变量 x_k 来表示题 6.3 的结果,其中,$x_k = x_0 + kh$.

解 ☞ 首先,注意到

$$k = \frac{x_k - x_0}{h}, \quad k - 1 = \frac{x_{k-1} - x_0}{h} = \frac{x_k - x_1}{h},$$

$$k - 2 = \frac{x_{k-2} - x_0}{h} = \frac{x_k - x_2}{h},$$

如此等等.以符号 $p(x_k)$ 代替 p_k,我们立即得到

$$p(x_k) = y_0 + \frac{\Delta y_0}{h}(x_k - x_0) + \frac{\Delta^2 y_0}{2!\,h^2}(x_k - x_0)(x_k - x_1)$$

$$+\cdots+\frac{\Delta^n y_0}{n!\,h^n}(x_k-x_0)\cdots(x_k-x_{n-1}).$$

这是另一形式的 Newton 公式.

6.5　找出一个 3 次多项式,该多项式所取的 4 个值排在对应于 x_k 列的 y_k 列中.

表 6.2

k	x_k	y_k	Δy_k	$\Delta^2 y_k$	$\Delta^3 y_k$
0	④	①			
			②		
1	⑥	3		③	
			5		④
2	⑧	8		7	
			12		
3	10	20			

解　☞ 所需的各种差分出现在表 6.2 的剩余列中.将带圈的数字代入 Newton 公式中相应的位置时,

$$p(x_k)=1+\frac{2}{2}(x_k-4)+\frac{3}{8}(x_k-4)(x_k-6)$$
$$+\frac{4}{48}(x_k-4)(x_k-6)(x_k-8).$$

它可被简化为

$$p(x_k)=\frac{1}{24}(2x_k^3-27x_k^2+142x_k-240),$$

不过,在应用中通常第一种形式更可取.

6.6　利用自变量 k 表示题 6.5 中的多项式.

解　☞ 直接由题 6.3,

$$p_k=1+2k+\frac{3}{2}k^{(2)}+\frac{4}{6}k^{(3)},$$

对于计算 p_k 值而言,它是方便的形式,因此可以原样保留下来.还能将它重新整理为

$$p_k=1+\frac{11}{6}k-\frac{1}{2}k^2+\frac{2}{3}k^3.$$

6.7　应用 Newton 公式,找出一个取表 6.3 的 y_k 值而不大于 4 次的多项式.

表 6.3

k	x_k	y_k	Δ	Δ^2	Δ^3	Δ^4
0	1	①				
			(-2)			
1	2	-1		④		
			2		(-8)	
2	3	1		-4		⑯
			(-2)		8	
3	4	1		4		
			2			
4	5	1				

解 ☞ 被圈起的是需要求的差分.将带圈的项代入 Newton 公式中相应的位置,

$$p_k = 1 - 2k + \frac{4}{2}k^{(2)} - \frac{8}{6}k^{(3)} + \frac{16}{24}k^{(4)},$$

它也是

$$p_k = \frac{1}{3}(2k^4 - 16k^3 + 40k^2 - 32k + 3).$$

由于 $k = x_k - 1$,故这个结果也能被写成

$$p(x_k) = \frac{1}{3}(2x_k^4 - 24x_k^3 + 100x_k^2 - 168x_k + 93).$$

补 充 题

6.8 找一个取下列值的 4 次多项式.

x_k	2	4	6	8	10
y_k	0	0	1	0	0

6.9 找一个取下列值的 2 次多项式.

$k = x_k$	0	1	2	3	4	5	6	7
y_k	1	2	4	7	11	16	22	29

6.10 找一个取下列值的 3 次多项式.

x_k	3	4	5	6
y_k	6	24	60	120

6.11 找一个取下列值的 5 次多项式.

$k = x_k$	0	1	2	3	4	5
y_k	0	0	1	1	0	0

6.12 找一个包含下列值的 3 次多项式(也见题 3.12).

$k = x_k$	0	1	2	3	4	5
y_k	1	2	4	8	15	26

6.13 以

$$p_k = a_0 + a_1k^{(1)} + a_2k^{(2)} + \cdots + a_nk^{(n)}$$

的形式,表示一个 n 次多项式,计算 $\Delta p_k, \Delta^2 p_k, \cdots, \Delta^n p_k$,然后试证:需要

$$p_k = y_k, \quad k = 0, \cdots, n,$$

才能导出 $\Delta p_0 = \Delta y_0, \Delta^2 p_0 = \Delta^2 y_0$,等等.其次,推出

$$a_0 = y_0, \quad a_1 = \Delta y_0, \quad a_2 = \frac{1}{2}\Delta^2 y_0, \cdots, a_n = \frac{1}{n!}\Delta^n y_0.$$

并将这些数代入,再一次得到 Newton 公式.

6.14 找一个在 $x = 0, 1, 2$ 上与 $y(x) = x^4$ 相配置的 2 次多项式.

6.15 找一个在 $x=0,1,2,3$ 上与 $y(x)=\sin\left(\frac{\pi x}{2}\right)$ 相配置的 3 次多项式,分别比较两个函数在 $x=4, x=5$ 时的值.

6.16 是否存在这样的一个 4 次多项式,它在 $x=0,1,2,3,4$ 上与 $y(x)=\sin\frac{\pi x}{2}$ 相配置?

6.17 是否存在这样的一个 2 次多项式,它在 $x=-1,0,1$ 上与 $y(x)=x^3$ 相配置?

6.18 找一个在 $x=-2,-1,0,1,2$ 上与 $y(x)=|x|$ 相配置的 4 次多项式.在何处该多项式比 $y(x)$ 大,何处又比 $y(x)$ 小?

6.19 找一个在 $x=0,1,4$ 上与 $y(x)=\sqrt{x}$ 相配置的 2 次多项式.为什么 Newton 公式不适用?

6.20 求 $\Delta^3 y_3=1$ 的一个解,它对于所有的 k 具有 $y_0=\Delta y_0=\Delta^2 y_0=0$.

第七章　算子与配置多项式

算子

在数值分析中到处都使用算子.特别,它适用于简化复杂公式的开发.某些最令人感兴趣的应用以乐观精神来实现,而不去特别拘泥于逻辑上的严密.其结果由其他方法来确认或用实验方法检验.

本章中导出的相当数量的公式具有某种程度上的历史影响,它们提供了早期的数值方面的具有优先权的见解,冠以像 Newton 与 Causs 这种名字,显示出他们在那个时期的重要性.计算硬件的变化已减少了它们的应用范围.第 12 章将重复要点,在那里将提供某些古典的应用.

立即要用到的特定的算子概念是这些:

1. **算子** Δ 由

$$\Delta y_k = y_{k+1} - y_k$$

定义.现在我们将 Δ 想象为这样一个运算:对于考虑中的所有 k 值,当提供 y_k 作为一个输入时,就产生 $y_{k+1} - y_k$ 作为一个输出.

$$\boxed{y_k} \longrightarrow \boxed{\Delta} \longrightarrow \boxed{y_{k+1} - y_k}$$

算子与一个算法(正如第一章中所叙述的)间的类似是显而易见的.

2. **算子** E 由

$$Ey_k = y_{k+1}$$

定义.这里,对于运算来说,输入仍是 y_k,而输出是 y_{k+1}.

$$\boxed{y_k} \longrightarrow \boxed{E} \longrightarrow \boxed{y_{k+1}}$$

Δ 和 E 均具有线性性质,即

$$\Delta(C_1 y_k + C_2 z_k) = C_1 \Delta y_k + C_2 \Delta z_k,$$
$$E(C_1 y_k + C_2 z_k) = C_1 E y_k + C_2 E z_k,$$

其中,C_1 与 C_2 是任意常数(与 k 无关).将要介绍的所有算子均具有这个性质.

3. **线性算子的组合**.考虑两个算子,分别记为 L_1, L_2,由输入 y_k,产生输出 $L_1 y_k$ 与 $L_2 y_k$,则这些算子的和定义为输出 $L_1 y_k + L_2 y_k$ 的运算.

$$\boxed{y_k} \longrightarrow \boxed{L_1 + L_2} \longrightarrow \boxed{L_1 y_k + L_2 y_k}$$

类似的定义引出两个算子的差.

更一般地.若 C_1 与 C_2 是常数(与 k 无关),则算子 $C_1 L_1 + C_2 L_2$ 产生输出 $C_1 L_1 y_k + C_2 L_2 y_k$.

$$\boxed{y_k} \longrightarrow \boxed{C_1 L_1 + C_2 L_2} \longrightarrow \boxed{C_1 L_1 y_k + C_2 L_2 y_k}$$

4. **算子 L_1 与 L_2 的积**定义为输出 $L_1 L_2 y_k$ 的运算.图表使之更清楚:

$$\boxed{y_k} \longrightarrow \boxed{L_2} \longrightarrow \boxed{L_2 y_k} \longrightarrow \boxed{L_1} \longrightarrow \boxed{L_1 L_2 y_k}$$

算子 L_1 作用于由 L_2 所产生的输出.图表中间的 3 个部分合在一起表示算子 $L_1 L_2$.

$$\boxed{y_k} \longrightarrow \boxed{L_1 L_2} \longrightarrow \boxed{L_1 L_2 y_k}$$

用这个积的定义,像上文中 C_1, C_2 这样的数也可看作算子.例如, C 是任意一数,算子 C 执行乘以数 C 的乘法运算.

5. **算子等式**.如果对所需考虑的问题中的所有输入,算子 L_1 与 L_2 产生完全相同的输出,则称它们相等.用符号表示即:对所需考虑的问题中的所有自变量 k,若

$$L_1 y_k = L_2 y_k,$$

则

$$L_1 = L_2.$$

由这个定义,对照输出立即看出:对任意的算子 L_1, L_2 与 L_3

$$L_1 + L_2 = L_2 + L_1,$$
$$L_1 + (L_2 + L_3) = (L_1 + L_2) + L_3,$$
$$L_1(L_2L_3) = (L_1L_2)L_3,$$
$$L_1(L_2 + L_3) = L_1L_2 + L_1L_3.$$

但乘法交换律不总是成立的:

$$L_1L_2 \neq L_2L_1.$$

然而,若两个算子中有一个是数 C,则由对照输出结果,显然有等式

$$CL_1 = L_2C.$$

6. **逆算子**.对于我们将使用的许多另外的算子来说,乘法交换律也成立.作为特例,若

$$L_1L_2 = L_2L_1 = 1$$

则称 L_1 与 L_2 为逆算子.在此情况下我们使用如下符号:

$$L_1 = L_2^{-1} = \frac{1}{L_2}, \quad L_2 = L_1^{-1} = \frac{1}{L_1}.$$

算子 1 就是通常所说的**恒等算子**.并且对任一算子,易见它使 $1 \cdot L = L \cdot 1$ 成立.

7. **连接 Δ 与 E 的简单方程**,其中包括

$$E = 1 + \Delta, \qquad \Delta^2 = E^2 - 2E + 1,$$
$$E\Delta = \Delta E, \qquad \Delta^3 = E^3 - 3E^2 + 3E - 1.$$

早期利用其他手段已证明了两个相关定理,用算子符号体系表示如下:

$$\Delta^k = \sum_{i=0}^{k}(-1)^i\binom{k}{i}E^{k-i}, \quad E^k = \sum_{i=0}^{k}\binom{k}{i}\Delta^i.$$

8. **向后差分算子**(backward difference operator) ∇ 由

$$\nabla y_k = y_k - y_{k-1}$$

定义.然后易证

$$\nabla E = E\nabla = \Delta.$$

∇ 与 E^{-1} 之间关系可证明是

$$E^{-1} = 1 - \nabla.$$

而对负整数 k,可导出展开式

$$y_k = y_0 + \sum_{i=1}^{k}\frac{k(k+1)\cdots(k+i-1)}{i!}\nabla^i y_0.$$

9. **中心差分算子**(central difference operator)由

$$\delta = E^{\frac{1}{2}} - E^{-\frac{1}{2}}$$

定义.接下来有 $\delta E^{\frac{1}{2}} = \Delta$.尽管带分数变元,但这是一个用得很多的算子.它与下面的算子有密切关系.

10. **平均算子**(averaging operator)由

$$\mu = \frac{1}{2}(E^{\frac{1}{2}} + E^{-\frac{1}{2}})$$

定义.从中心差分运算中消去分数变元是它的主要作用.

配置多项式

至此,配置多项式能表示为多种可供选择的形式,它们均等价于第 6 章的 Newton 公式,而每一个都适合于多少有些不同的情况.我们讨论如下公式,它们从第十二章开始找到用途.

1. **Newton 向后公式**(Newton's backward formula)

$$p_k = y_0 + k\nabla y_0 + \frac{k(k+1)}{2!}\nabla^2 y_0 + \cdots + \frac{k\cdots(k+n-1)}{n!}\nabla^n y_0$$

表示当 $k=0,-1,\cdots,-n$ 时取值为 y_k 的配置多项式.

2. **Gauss 向前公式**(Gauss forward formula)可通过展开 E 与 δ 间的关系得到.若多项式具有偶数 $2n$ 次且配置在 $k=-n,\cdots,n$ 上,则为

$$p_k = y_0 + \sum_{i=1}^{n}\left[\binom{k+i-1}{2i-1}\delta^{2i-1}y_{1/2} + \binom{k+i-1}{2i}\delta^{2i}y_0\right];$$

若多项式具有奇数 $2n+1$ 次且配置在 $k=-n,\cdots,n+1$ 上,则变为

$$p_k = \sum_{i=0}^{n}\left[\binom{k+i-1}{2i}\delta^{2i}y_0 + \binom{k+1}{2i+1}\delta^{2i+1}y_{1/2}\right].$$

3. **Gauss 向后公式**(Gauss backward formula)能用类似的方法获得.对于偶数次它仍取配置在 $k=-n,\cdots,n$ 上的形式

$$p_k = y_0 + \sum_{i=1}^{n}\left[\binom{k+i-1}{2i-1}\delta^{2i-1}y_{-\frac{1}{2}} + \binom{k+i}{2i}\delta^{2i}y_0\right].$$

这两个 Gauss 公式的一个重要用途是推出 Stirling 公式.

4. **Stirling 公式**是配置多项式应用得最多的形式之一.它形如

$$p_k = y_0 + \binom{k}{1}\delta\mu y_0 + \left(\frac{k}{2}\right)\binom{k}{1}\delta^2 y_0 + \binom{k+1}{3}\delta^3\mu y_0$$
$$+\left(\frac{k}{4}\right)\binom{k+1}{3}\delta^4 y_0 + \cdots + \binom{k+n-1}{2n-1}\delta^{2n-1}\mu y_0 + \left(\frac{k}{2n}\right)\binom{k+n-1}{2n-1}\delta^{2n}y_0,$$

并且是一个很受欢迎的公式.不言而喻.它配置在 $k=-n,\cdots,n$ 上.

5. **Everett 公式**取如下形式:

$$p_k = \binom{k}{1}y_1 + \binom{k+1}{3}\delta^2 y_1 + \binom{k+2}{5}\delta^4 y_1 + \cdots + \binom{k+n}{2n+1}\delta^{2n}y_1$$
$$-\binom{k-1}{1}y_0 - \binom{k}{3}\delta^2 y_0 - \binom{k+1}{5}\delta^4 y_0 - \cdots - \binom{k+n-1}{2n+1}\delta^{2n}y_0,$$

并且可由重排奇数次 Gauss 向前公式的构成成分来获得,它配置在 $k=-n,\cdots,n+1$ 上.注意,式中仅出现偶阶差分.

6. **Bessel 公式**是 Everett 公式的一个重排,并且能写作

$$p_k = \mu y_{\frac{1}{2}} + \left(k-\frac{1}{2}\right)\delta y_{\frac{1}{2}} + \binom{k}{2}\mu\delta^2 y_{\frac{1}{2}} + \frac{1}{3}\left(k-\frac{1}{2}\right)\binom{k}{2}\delta^3 y_{\frac{1}{2}}$$
$$+\cdots+\binom{k+n-1}{2n}\mu\delta^{2n}y_{\frac{1}{2}} + \left(\frac{1}{2n+1}\right)\left(k-\frac{1}{2}\right)\binom{k+n-1}{2n}\delta^{2n+1}y_{\frac{1}{2}}.$$

题　解

7.1　证明 $E=1+\Delta$.

证　☞ 根据 E 的定义,$Ey_k=y_{k+1}$;而根据 $1+\Delta$ 的定义,

$$(1+\Delta)=1\cdot y_k+\Delta y_k = y_k+(y_{k+1}-y_k)=y_{k+1}.$$

对于所有的自变量 k,算子 E 与 $1+\Delta$ 具有完全一样的输出,故两者相等.这个结果也可写成 $\Delta=E-1$.

7.2 证明 $E\Delta=\Delta E$.

证 ☞ 因

$$E\Delta y_k=E(y_{k+1}-y_k)=y_{k+2}-y_{k+1},$$

且

$$\Delta E y_k=\Delta y_{k+1}=y_{k+2}-y_{k+1},$$

输出相等,从而算子相等.这是一个乘法交换律成立的例子.

7.3 证明 $\Delta^2=E^2-2E+1$.

证 ☞ 利用算子的多种性质,

$$\Delta^2=(E-1)(E-1)=E^2-1\cdot E-E\cdot 1+1=E^2-2E+1.$$

7.4 应用二项式定理证明 $\Delta^k y_0=\sum_{i=0}^{k}(-1)^i\binom{k}{i}y_{k-i}$.

证 ☞ 只要 a 与 b(从而与 $a+b$)在乘法运算中可交换,则二项式定理

$$(a+b)^k=\sum_{i=0}^{k}\binom{k}{i}a^{k-i}b^i$$

就是正确的.在当前情况下,这些元素将是 E 与 -1,而它们可交换,故

$$\Delta^k=(E-1)^k=\sum_{i=0}^{k}(-1)^i\binom{k}{i}E^{k-i}.$$

注意到 $Ey_0=y_1$, $E^2y_0=y_2$,等等,最后我们有

$$\Delta^k y_0=\sum_{i=0}^{k}(-1)^i\binom{k}{i}y_{k-i}.$$

这重复了题 3.5 的结果.

7.5 证明 $y_k=\sum_{i=0}^{k}\binom{k}{i}\Delta^i y_0$.

证 ☞ 由于 $E=1+\Delta$,故由二项式定理得到

$$E^k=(1+\Delta)^k=\sum_{i=0}^{k}\binom{k}{i}\Delta^i.$$

对 y_0 应用这个算子,并且利用 $E^k y_0=y_k$ 这一事实,立即得到所需结果.注意,这重复了题 6.2.

7.6 **向后差分**(backward difference)由

$$\nabla y_k=y_k-y_{k-1}=\Delta y_{k-1}$$

所定义.显然,它包含着对 y_k-y_{k-1} 指定一个新的符号.试证 $\nabla E=E\nabla=\Delta$, $E^{-1}=1-\nabla$.

证 ☞ 由于对所有的自变量 k,均有

$$\nabla E y_k=\nabla y_{k+1}=y_{k+1}-y_k=\Delta y_k,$$
$$E\nabla y_k=E(y_k-y_{k-1})=y_{k+1}-y_k=\Delta y_k,$$

所以我们有

$$\nabla E=E\nabla=\Delta=E-1.$$

对于由 $E^{-1}y_k=y_{k-1}$ 所定义的算子,利用符号 E^{-1},我们发现 $EE^{-1}y_k$ 与 $E^{-1}\cdot Ey_k$ 都是 y_k.用算子语言来说,意味着这两个算子是互逆的:$EE^{-1}=E^{-1}E=1$.最后,作为算子计算的一个练习,

$$\nabla=E^{-1}E\nabla=E^{-1}\Delta=E^{-1}(E-1)=1-E^{-1},$$

而

$$E^{-1}=1-\nabla.$$

7.7 如表 7.1 所示,利用负的变量 k,向后差分通常仅用在一个表的底部.利用符号 $\nabla^2 y_k=\nabla\nabla y_k$, $\nabla^3 y_k=\nabla\nabla^2 y_k$ 等,试证 $\Delta^n y_k=\nabla^n y_{k+n}$.

表 7.1

k	x	y				
-4	x_{-4}	y_{-4}				
			∇y_{-3}			
-3	x_{-3}	y_{-3}		$\nabla^2 y_{-2}$		
			∇y_{-2}		$\nabla^3 y_{-1}$	
2	x_{-2}	y_{-2}		$\nabla^2 y_{-1}$		$\nabla^4 y_0$
			∇y_{-1}		$\nabla^3 y_0$	
1	x_{-1}	y_{-1}		$\nabla^2 y_0$		
			∇y_0			
0	x_0	y_0				

证 ☞ 由于 $\Delta = E\nabla$，我们有 $\Delta^n = (E\nabla)^n$，而 E 与 ∇ 可交换，于是式子右端的 $2n$ 个因子重新整理后可给出 $\Delta^n = \nabla^n E^n$。将它作用于 y_k，$\Delta^n y_k = \nabla^n E^n y_k = \nabla^n y_{k+n}$.

7.8 证明

$$y_{-1} = y_0 - \nabla y_0, \quad y_{-2} = y_0 - 2\nabla y_0 + \nabla^2 y_0,$$
$$y_{-3} = y_0 - 3\nabla y_0 + 3\nabla^2 y_0 - \nabla^3 y_0.$$

并证明，对于负整数 k，一般有

$$y_k = y_0 + \sum_{i=1}^{-k} \frac{k(k+1)\cdots(k+i-1)}{i!} \nabla^i y_0.$$

证 ☞ 按照一般情况，立即有

$$y_k = E^k y_0 = (E^{-1})^{-k} y_0 = (1-\nabla)^{-k} y_0.$$

利用带有负整数 k 的二项式定理，使得

$$y_k = \sum_{i=0}^{k} (-1)^i \binom{-k}{i} \nabla^i y_0 = y_0 + \sum_{i=1}^{-k} (-1)^i \frac{(-k)(-k-1)\cdots(-k-i+1)}{i!} \nabla^i y_0$$
$$= y_0 + \sum_{i=1}^{-k} \frac{k(k+1)\cdots(k+i-1)}{i!} \nabla^i y_0.$$

而对 $k=-1,-2,-3$ 这几种特殊情况，写出和式后便可立得.

7.9 证明：具有由下面的公式

$$p_k = y_0 + k\nabla y_0 + \frac{k(k+1)}{2!}\nabla^2 y_0 + \cdots + \frac{k\cdots(k+n-1)}{n!}\nabla^n y_0$$
$$= y_0 + \sum_{i=1}^{n} \frac{k(k+1)\cdots(k+i-1)}{i!} \nabla^i y_0$$

定义其值的 n 次多项式，当 $k=0,-1,\cdots,-n$ 时，可简化为 $p_k = y_k$（这是向后 Newton 差分公式）.

证 ☞ 该证明与题 6.3 中的证明非常相像. 当 $k=0$ 时，式子的右端仅剩下第一项；当 $k=-1$ 时，仅剩前两项，其余的项均为零. 一般地，若 k 是从 0 到 $-n$ 的任一整数，则对于 $i>-k$，$k(k+1)\cdots(k+i-1)=0$，和式可缩写为

$$p_k = y_0 + \sum_{i=1}^{-k} \frac{k(k+1)\cdots(k+i-1)}{i!} \nabla^i y_0,$$

而由题 7.8，这可简化为 y_k. 因而，对于 $k=0,-1,\cdots,-n$，这个问题中的多项式与我们的 y_k 函数一致.

7.10 找出一个 3 次多项式，该多项式所取的 4 个值列在表 7.2 对应 x_k 的 y_k 处.

表 7.2

k	x_k	y_k	∇y_k	$\nabla^2 y_k$	$\nabla^3 y_k$
−3	4	1			
			2		
2	6	3		3	
			5		①
1	8	8		○	
			⑫		
0	10	⑳			

解 ☞ 所需求出的差分出现在表 7.2 余下的列中,将带圈的数字代入 Newton 向后公式相应的位置中,

$$p_k = 20 + 12k + \frac{7}{2}k(k+1) + \frac{4}{6}k(k+1)(k+2).$$

注意到除了自变量 k,这组数据与题 6.5 中的相同.由关系式 $x_k = 10 + 2k$ 消去 k,在题 6.5 中得到的公式

$$p(x_k) = \frac{1}{24}(2x_k^3 - 27x_k^2 + 142x_k - 240)$$

再一次被得到.两个 Newton 公式不过是相同的多项式的重排.随后还有其他的重排.

7.11 中心差分算子 δ 由 $\delta = E^{\frac{1}{2}} - E^{-\frac{1}{2}}$ 所定义,从而 $\delta y_{\frac{1}{2}} = y_1 - y_0 = \Delta y_0 = \nabla y_1$,如此等等.注意到 $E^{\frac{1}{2}}$ 与 $E^{-\frac{1}{2}}$ 是互逆的,且 $(E^{\frac{1}{2}})^2 = E$,$(E^{-\frac{1}{2}})^2 = E^{-1}$,试证 $\Delta^n y_k = \delta^n y_{k+\frac{n}{2}}$.

证 ☞ 根据 δ 的定义,我们有 $\delta E^{\frac{1}{2}} = E - 1 = \Delta$ 与 $\Delta^n = \delta^n E^{\frac{n}{2}}$,将它应用于 y_k,产生所需的结果.

7.12 用 δ 符号,通常的差分表可改写为表 7.3.

表 7.3

k	y_k	δ	δ^2	δ^3	δ^4
−2	y_{-2}				
		$\delta y_{-\frac{3}{2}}$			
−1	y_{-1}		$\delta^2 y_{-1}$		
		$\delta y_{-\frac{1}{2}}$		$\delta^3 y_{-\frac{1}{2}}$	
0	y_0		$\delta^2 y_0$		$\delta^4 y_0$
		$\delta y_{\frac{1}{2}}$		$\delta^3 y_{\frac{1}{2}}$	
1	y_1		$\delta^2 y_1$		
		$\delta y_{\frac{3}{2}}$			
2	y_2				

利用 Δ 算子表示 $\delta y_{\frac{1}{2}}$,$\delta^2 y_0$,$\delta^3 y_{\frac{1}{2}}$ 与 $\delta^4 y_0$.

解 ☞ 根据题 7.11,$\delta y_{\frac{1}{2}} = \Delta y_0$,$\delta^2 y_0 = \Delta^2 y_{-1}$,$\delta^3 y_{\frac{1}{2}} = \Delta^3 y_{-1}$,$\delta^4 y_0 = \Delta^4 y_{-2}$.

7.13 **平均算子** μ 由 $\mu = \frac{1}{2}(E^{\frac{1}{2}} + E^{-\frac{1}{2}})$ 定义,从而 $\mu y_{\frac{1}{2}} = \frac{1}{2}(y_1 + y_0)$,等等.证明 $\mu^2 = 1 + \frac{1}{4}\delta^2$.

证 ☞ 我们首先计算 $\delta^2 = E - 2 + E^{-1}$,然后

$$\mu^2 = \frac{1}{4}(E + 2 + E^{-1}) = \frac{1}{4}(\delta^2 + 4) = 1 + \frac{1}{4}\delta^2.$$

7.14 对于指定的自变量 k,验证如下各项

$k = 0, 1,\qquad y_k = y_0 + \binom{k}{1}\delta y_{\frac{1}{2}}$;

$k=-1,0,1,\qquad y_k=y_0+\binom{k}{1}\delta y_{\frac12}+\binom{k}{2}\delta^2 y_0;$

$k=-1,0,1,2,\qquad y_k=y_0+\binom{k}{1}\delta y_{\frac12}+\binom{k}{2}\delta^2 y_0+\binom{k+1}{2}\delta^3 y_{\frac12};$

$k=-2,-1,0,1,2,\qquad y_k=y_0+\binom{k}{1}\delta y_{\frac12}+\binom{k}{2}\delta^2 y_0+\binom{k+1}{3}\delta^3 y_{\frac12}+\binom{k+1}{4}\delta^4 y_0.$

证 ☞ 当 $k=0$ 时,式子的右端仅有 y_0 项.当 $k=1$ 时,所有的右端对应于算子

$$1+\delta E^{\frac12}=1+(E-1)=E,$$

其结果产生 y_1.当 $k=-1$ 时,后三个公式导出

$$1-\delta E^{\frac12}+\delta^2=1-(E-1)+(E-2+E^{-1})=E^{-1},$$

它产生 y_{-1}.当 $k=2$ 时,后两个公式导致

$$1+2\delta E^{\frac12}+\delta^2+\delta^3E^{\frac12}=1+2(E-1)+(E-2+E^{-1})(1+E-1)=E^2,$$

产生出 y_2.最后,当 $k=-2$ 时,最后的公式包含

$$1-2\delta E^{\frac12}+3\delta^2-\delta^3E^{\frac12}+\delta^4=1-2(E-1)+(E-2+E^{-1})[3-(E-1)-(E-2+E^{-1})]=E^{-2},$$

它导出 y_{-2}.

推广这个问题的公式形成 **Gauss 向前公式**.它表示一个对于 $k=-n,\cdots,n$,取值 $p_k=y_k$ 的 $2n$ 次多项式

$$p_k=y_0+\sum_{i=1}^{n}\left[\binom{k+i-1}{2i-1}\delta^{2i-1}y_{\frac12}+\binom{k+i-1}{2i}\delta^{2i}y_0\right],$$

或表示一个对于 $k=-n,\cdots,n+1$,取值 $p_k=y_k$ 的 $2n+1$ 次多项式

$$p_k=\sum_{i=0}^{n}\left[\binom{k+i-1}{2i}\delta^{2i}y_0+\binom{k+i}{2i+1}\delta^{2i+1}y_{\frac12}\right]$$

(在特殊情况下,次数可能较低).

7.15 利用 $n=2$ 的 Gauss 公式找出一个在表 7.4 中取 y_k 值的、不大于 4 次的多项式.

表 7.4

k	x_k	y_k				
−2	2	−2				
			3			
−1	4	1		1		
			2		4	
0	6	③		③		⓪
			⑤		④	
1	8	8		7		
			12			
2	10	20				

解 ☞ 所需的差分通常被编制成表.这类似于在阐述两个 Newton 公式中所用的一个函数,但这里,自变量 k 带一个位移,并且在顶部增加一个额外的数对.由于此例中 4 阶差分是零,故我们预料它是一个 3 次多项式.将圈起的数字代入 Gauss 公式中相应的位置,

$$p_k=3+5k+\frac32k(k-1)+\frac46(k+1)k(k-1).$$

若由关系式 $x_k=6+2k$ 消去 k,则已经找出过两次的 3 次式将又一次出现.

7.16 应用 Gauss 向前公式找出一个在表 7.5 中取 y_k 值的不大于 4 次的多项式.

表 7.5

k	x_k	y_k				
-2	1	1				
			2			
-1	2	-1		4		
			2		-8	
0	3	①		$-$④		⑯
			$-$②		⑧	
1	4	-1		4		
			2			
2	5	1				

解　☞ 将所需的差分圈起，代入 Gauss 公式的相应位置，

$$p_k = 1 - 2k - 4\frac{k(k-1)}{2} + 8\frac{(k+1)k(k-1)}{6} + 16\frac{(k+1)k(k-1)(k-2)}{24}.$$

它可简化为

$$p_k = \frac{1}{3}(2k^4 - 8k^2 + 3).$$

由于 $k = x_k - 3$，于是这一结果也可写为

$$p(x_k) = \frac{1}{3}(2x_k^4 - 24x_k^3 + 100x_k^2 - 168x_k + 93).$$

显然，这与以前由 Newton 公式求出的多项式一致.

7.17　对于 $k = -1, 0, 1$，验证

$$y_k = y_0 + \binom{k}{1}\delta y_{-\frac{1}{2}} + \binom{k+1}{2}\delta^2 y_0;$$

然后，对于 $k = -2, -1, 0, 1, 2$，验证

$$y_k = y_0 + \binom{k}{1}\delta y_{-\frac{1}{2}} + \binom{k+1}{2}\delta^2 y_0 + \binom{k+1}{3}\delta^3 y_{-\frac{1}{2}} + \binom{k+2}{4}\delta^4 y_0.$$

证　☞ 对于 $k = 0$，式子右端仅有 y_0 项. 当 $k = 1$ 时，两式均含算子

$$1 + \delta E^{-\frac{1}{2}} + \delta^2 = 1 + (1 - E^{-1}) + (E - 2 + E^{-1}) = E,$$

它确实产生 y_1. 对于 $k = -1$，两式均含

$$1 - \delta E^{-\frac{1}{2}} = 1 - (1 - E^{-1}) = E^{-1},$$

它确实产生 y_{-1}. 继续使用第二个公式，对于 $k = 2$，我们有

$$\begin{aligned}&1 + 2\delta E^{-\frac{1}{2}} + 3\delta^2 + \delta^3 E^{-\frac{1}{2}} + \delta^4\\ = &1 + 2(1 - E^{-1}) + (E - 2 + E^{-1})(3 + 1 - E^{-1} + E - 2 + E^{-1}) = E^2.\end{aligned}$$

并且当 $k = -2$ 时，

$$\begin{aligned}1 - 2\delta E^{-\frac{1}{2}} + \delta^2 - \delta^3 E^{-\frac{1}{2}} &= 1 - 2(1 - E^{-1}) + (E - 2 + E^{-1})(1 - 1 + E^{-1})\\ &= E^{-2}\end{aligned}$$

正是所要求的.

此问题的公式可推广而形成 Gauss 向后公式. 它表示与偶数阶 Gauss 向前公式相同的多项式，并且如上所述，能验证

$$p_k = y_0 + \sum_{i=1}^{n}\left[\binom{k+i-1}{2i-1}\delta^{2i-1}y_{-\frac{1}{2}} + \binom{k+i}{2i}\delta^{2i}y_0\right].$$

7.18　证明 $\binom{k+i}{2i} + \binom{k+i-1}{2i} = \frac{k}{i}\binom{k+i-1}{2i-1}$.

证 ☞ 由二项式系数的定义，

$$\binom{k+i}{2i}+\binom{k+i-1}{2i}=\binom{k+i-1}{2i-1}[(k+i)+(k-i)]\frac{1}{2i}$$

正是所要求的.

7.19 由 Gauss 公式，推出下面给出的 Stirling 公式.

解 ☞ 将 $2n$ 次的 Gauss 公式逐项相加，除以 2，然后利用题 7.18 得

$$p_k=y_0+\sum_{i=1}^{n}\left[\binom{k+i-1}{2i-1}\delta^{2i-1}\mu y_0+\frac{k}{2i}\binom{k+i-1}{2i-1}\delta^{2i}y_0\right]$$

$$=y_0+\binom{k}{1}\delta\mu y_0+\frac{k}{2}\binom{k}{1}\delta^2 y_0+\binom{k+1}{3}\delta^3\mu y_0+\frac{4}{k}\binom{k-1}{3}\delta^4 y_0$$

$$+\cdots+\binom{k+n-1}{2n-1}\delta^{2n-1}\mu y_0+\frac{k}{2n}\binom{k+n-1}{2n-1}\delta^{2n}y_0.$$

这就是 Stirling 公式.

7.20 应用 $n=2$ 的 Stirling 公式求一个在表 7.6 中取 y_k 值的不大于 4 次的多项式.

表 7.6

k	x_k	y_k	δ	δ^2	δ^3	δ^4
−2	2	−2				
			3			
−1	4	1		−1		
			②		④	
0	6	③		③		⓪
			⑤		④	
1	8	8		7		
			12			
2	10	20				

解 ☞ 仍列出所需的差分表.将圈起的项代入 Stirling 公式中相应的位置，

$$p_k=3+\frac{2+5}{2}k+3\frac{k^2}{2}+\frac{4+4}{2}\frac{(k+1)k(k-1)}{6},$$

易见，这是由 Gauss 向前公式所得结果的一个较小的重排.

7.21 证明

$$\binom{k+i-1}{2i}\delta^{2i}y_0+\binom{k+i}{2i+1}\delta^{2i+1}y_{\frac{1}{2}}=\binom{k+i}{2i+1}\delta^{2i}y_1-\binom{k+i-1}{2i+1}\delta^{2i}y_0.$$

证 ☞ 左边变为(利用题 4.5)

$$\left[\binom{k+i}{2i+1}-\binom{k+i-1}{2i+1}\right]\delta^{2i}y_0+\binom{k+i}{2i+1}\delta^{2i+1}y_{\frac{1}{2}}$$

$$=\binom{k+i}{2i+1}[\delta^{2i}(1+\delta E^{\frac{1}{2}})y_0]-\binom{k+i-1}{2i+1}\delta^{2i}y_0$$

$$=\binom{k+i}{2i+1}\delta^{2i}y_1-\binom{k+i-1}{2i+1}\delta^{2i}y_0.$$

其中在最后一步中，我们使用了 $1+\delta E^{\frac{1}{2}}=E$.

7.22 试由奇数阶 Gauss 向前公式推导出 Everett 公式.

解 ☞ 利用题 7.1，我们立即得到

$$p_k=\sum_{i=0}^{n}\left[\binom{k+i}{2i+1}\delta^{2i}y_1-\binom{k+i-1}{2i+1}\delta^{2i}y_0\right]$$

$$=\binom{k}{1}y_1+\binom{k+1}{3}\delta^2y_1+\binom{k+2}{5}\delta^4y_1+\cdots+\binom{k+n}{2n+1}\delta^{2n}y_1$$

$$-\binom{k-1}{1}y_0-\binom{k}{3}\delta^2y_0-\binom{k+1}{5}\delta^4y_0-\cdots-\binom{k+n-1}{2n-1}\delta^{2n}y_0.$$

此即 Everett 公式.由于它是 Gauss 公式的一个重排,故对于 $k=-n,\cdots,n+1$,它是满足 $p_k=y_k$ 的同样的 $2n+1$ 次多项式.因为它的简单性,仅含偶数阶差分,所以它是一个用得很多的公式.

7.23 应用 $n=2$ 的 Everett 公式,求出一个取表 7.7 的 y_k 值的不大于 4 次的多项式.

表 7.7

k	x_k	y_k	δ	δ^2	δ^3	δ^4
−2	0	0				
			−1			
−1	1	−1		10		
			9		108	
0	2	(8)		(118)		(216)
			127		324	
1	3	(135)		(442)		(336)
			569		660	
2	4	704		1102		
			1671			
3	5	2375				

解 ☞ 所需的差分被圈起.

将带圈的项代入 Everett 公式中相应位置时,

$$p_k=135k+442\frac{(k+1)k(k-1)}{6}+336\frac{(k+2)(k+1)k(k-1)(k-2)}{120}$$

$$-8(k-1)-118\frac{k(k-1)(k-2)}{6}-216\frac{(k+1)k(k-1)(k-2)(k-3)}{120},$$

利用 $x_k=k+2$,可将其简化为

$$p(x_k)=x_k^5-x_k^4-x_k^3.$$

7.24 试证

$$\binom{k+i-1}{2i}\mu\delta^{2i}y_{\frac12}+\frac{k-\frac12}{2i+1}\binom{k+i-1}{2i}\delta^{2i+1}y_{\frac12}$$

$$=\binom{k+i}{2i+1}\delta^{2i}y_1-\binom{k+i-1}{2i+1}\delta^{2i}y_0.$$

证 ☞ 上式左端对应于算子

$$\delta^{2i}\binom{k+i-1}{2i}\frac{1}{2}\left[E+1+\frac{2k-1}{2i+1}(E-1)\right]$$

$$=\delta^{2i}\binom{k+i-1}{2i}\left(\frac{k+i}{2i+1}E-\frac{k-i-1}{2i+1}\right);$$

其右端对应于算子

$$\delta^{2i}\left[\binom{k+i}{2i+1}E-\binom{k+i-1}{2i+1}\right]=\delta^{2i}\binom{k+i-1}{2i}\left(\frac{k+i}{2i+1}E-\frac{k-i-1}{2i+1}\right),$$

从而两端相等.

7.25　试证 Bessel 公式是 Everett 公式的一个重排.

证　☞ Bessel 公式是

$$p_k=\sum_{i=0}^{n}\left[\binom{k+i-1}{2i}\mu\delta^{2i}y_{\frac{1}{2}}+\frac{1}{2i+1}\left(k-\frac{1}{2}\right)\binom{k+i-1}{2i}\delta^{2i+1}y_{\frac{1}{2}}\right]$$

$$=\mu y_{\frac{1}{2}}+\left(k-\frac{1}{2}\right)\delta y_{\frac{1}{2}}+\binom{k}{2}\mu\delta^2 y_{\frac{1}{2}}+\frac{1}{3}\left(k-\frac{1}{2}\right)\binom{k}{2}\delta^3 y_{\frac{1}{2}}$$

$$+\cdots+\binom{k+n-1}{2n}\mu\delta^{2n}y_{\frac{1}{2}}+\frac{1}{2n+1}\left(k-\frac{1}{2}\right)\binom{k+n-1}{2n}\delta^{2n+1}y_{\frac{1}{2}}.$$

根据前题,可立即简化为 Everett 公式.

7.26　使用 $n=1$ 的 Bessel 公式,求出一个取表 7.8 的 y_k 值的不大于 3 次的多项式.

表 7.8

k	x_k	y_k			
−1	4	1			
			2		
0	6	③		③	
			⑤		①
1	8	⑧		⑦	
			12		
2	10	20			

解　☞ 将所需的差分圈起来,并且将它们插入 Bessel 公式中相应的位置,无需置疑,产生的多项式与已由其他公式获得的多项式相同,

$$p_k=\frac{3+8}{2}+5\left(k-\frac{1}{2}\right)+\frac{3+7}{2}\,\frac{k(k-1)}{2}+\frac{1}{3}(4)\left(k-\frac{1}{2}\right)\frac{k(k-1)}{2}.$$

可以验证,它与早先的结果是等价的.

补　充　题

7.27　证明 $\nabla=\delta E^{-\frac{1}{2}}=1-E^{-1}=1-(1+\Delta)^{-1}$.

7.28　证明 $\sqrt{1+\delta^2\mu^2}=1+\frac{1}{2}\delta^2$.

7.29　证明 $E^{\frac{1}{2}}=\mu+\frac{1}{2}\delta$ 且 $E^{-\frac{1}{2}}=\mu-\frac{1}{2}\delta$.

7.30　若 $L_1L_2=L_2L_1$,两个算子 L_1 与 L_2 可交换.试证 μ,δ,E,Δ 与 ∇ 相互间均可交换.

7.31　证明 $\mu\delta=\frac{1}{2}\Delta E^{-1}+\frac{1}{2}\Delta$.

7.32　证明 $\Delta=\frac{1}{2}\delta^2+\delta\sqrt{1+\frac{1}{4}\delta^2}$.

7.33　将 Newton 向后公式用于以下数据,以获得自变量 k 的一个 4 次多项式.然后用 $x_k=k+5$ 将其转换成 x_k 的一个多项式,将最后结果与题 6.7 的结果相比较.

k	−4	−3	−2	−1	0
x_k	1	2	3	4	5
y_k	1	−1	1	−1	1

7.34 利用 Newton 向后公式求一个 3 次多项式，它含如下 x_k, y_k 数对：

x_k	3	4	5	6
y_k	6	24	60	120

用 $x_k = k + 6$ 将它转换成 x_k 的多项式，并且与题 6.10 的结果相比较.

7.35 试证自变量 $x_k = x_0 + kh$ 这一变化能将 Newton 向后公式转换为

$$p(x_k) = y_0 + \frac{\nabla y_0}{h}(x - x_0) + \frac{\nabla^2 y_0}{2!\,h^2}(x - x_0)(x - x_{-1}) + \cdots + \frac{\nabla^n y_0}{n!\,h^n}(x - x_0)\cdots(x - x_{-n+1}).$$

7.36 对于题 7.34 的数据，利用题 7.35 去直接获得自变量 x_k 的 3 次多项式.

7.37 将 Gauss 向前公式用于下面的数据，并且将结果与题 6.8 的结果相比较.

k	−2	−1	0	1	2
x_k	2	4	6	8	10
y_k	0	0	1	0	0

7.38 将 Gauss 向后公式用于题 7.37 的数据.

7.39 将 Gauss 向后公式用于题 7.34 的数据，其中，自变量 k 位移，使得 $x = 6$ 时 $k = 0$.

7.40 将 Gauss 向前公式用于下面的数据，并与题 6.11 的结果相比较.

k	−2	−1	0	1	2	3
x_k	0	1	2	3	4	5
y_k	0	0	1	1	0	0

7.41 对于 $k = -1, 0$，验证

$$y_k = y_0 + \binom{k}{1}\delta y_{-\frac{1}{2}}.$$

而对于 $k = -2, -1, 0, 1$，

$$y_k = y_0 + \binom{k}{1}\delta y_{-\frac{1}{2}} + \binom{k+1}{2}\delta^2 y_0 + \binom{k+1}{3}\delta^3 y_{-\frac{1}{2}}.$$

这些也能被看成是 Gauss 向后公式的形式，这些多项式的次数是奇数而非偶数.

7.42 将 Stirling 公式用于题 7.37 的数据.

7.43 将 Stirling 公式用于题 6.9 的数据. 选择任意 3 个等距自变量且设它们对应于 $k = -1, 0, 1$.

7.44 将 Everett 公式用于题 7.34 的数据，取对应于 $k = 0$ 与 1 的自变量的中心数对.

7.45 将 Everett 公式用于题 7.40 的数据.

7.46 将 Everett 公式用于题 6.9 的数据.

7.47 将 Bessel 公式用于题 7.44 的数据.

7.48 将 Bessel 公式用于题 7.40 的数据.

7.49 证明 $E^{\frac{1}{2}} = \frac{1}{2}\delta + \mu = \left(1 + \frac{1}{4}\delta^2\right)^{\frac{1}{2}} + \frac{1}{2}\delta = 1 + \frac{1}{2}\delta + \frac{1}{8}\delta^2 + \cdots$.

7.50 试证 $\mu^{-1} = 1 - \frac{1}{8}\delta^2 + \frac{3}{128}\delta^4 - \frac{5}{1024}\delta^6 + \cdots$.

7.51 证明 $\delta(f_k g_k) = \mu f_k \delta g_k + \mu g_k \delta f_k$.

第八章　不等距自变量

对于不等距自变量 $x_0, \cdots, x_n$，能用多种方法找到配置多项式. 本章将介绍 Lagrange 法、行列式法和均差(divided difference)法.

1. **Lagrange 公式**是

$$p(x) = \sum_{i=0}^{n} L_i(x) y_i,$$

其中，$L_i(x)$ 是 Lagrange 乘子函数(Lagrange multiplier function)：

$$L_i(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)},$$

它具有如下性质：

$$L_i(x_k) = 0, \quad k \neq i; L_i(x_i) = 1.$$

Lagrange 公式的确表示了配置多项式，即，对 $k=0, \cdots, n$，$p(x_k) = y_k$. 函数

$$\pi(x) = (x-x_0)\cdots(x-x_n) = \prod_{i=0}^{n}(x-x_i)$$

能用来以更紧凑的形式表示 Largrange 乘子函数

$$L_i(x) - \frac{\pi(x)}{(x-x_i)\pi'(x_i)}.$$

密切相关的函数

$$F_k(x) = \prod_{i \neq k}(x-x_i)$$

导致 Lagrange 乘子函数的第二个紧凑表示法

$$L_i(x) = \frac{F_i(x)}{F_i(x_i)}.$$

2. 配置多项式 $p(x)$ 的**行列式形式**是

$$\begin{vmatrix} p(x) & 1 & x & x^2 & \cdots & x^n \\ y_0 & 1 & x_0 & x_0^2 & \cdots & x_0^n \\ y_1 & 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y_n & 1 & x_n & x_n^2 & \cdots & x_n^n \end{vmatrix} = 0,$$

这因为当 $k=0, \cdots, n$ 时，$p(x_k) = y_k$. 它偶尔有用，主要用于理论工作.

3. x_0 与 x_1 之间的**第一均差**由

$$y(x_0, x_1) = \frac{y_1 - y_0}{x_1 - x_0}$$

所定义. 在其他的自变量数对间可应用类似的公式.

于是，**较高阶的均差**可借助于较低阶的均差来定义，例如

$$y(x_0, x_1, x_2) = \frac{y(x_1, x_2) - y(x_0, x_1)}{x_2 - x_0}$$

是一个 2 阶均差，而

$$y(x_0, x_1, \cdots, x_n) = \frac{y(x_1, \cdots, x_n) - y(x_0, \cdots, x_{n-1})}{x_n - x_0}$$

是一个 n 阶均差. 在许多地方，这些差分扮演着与以前使用过的较简单的差分等价的

角色.

差分表仍是展示差分的方便工具,用的是标准的对角线形式

x_0	y_0				
		$y(x_0,x_1)$			
x_1	y_1		$y(x_0,x_1,x_2)$		
		$y(x_1,x_2)$		$y(x_0,x_1,x_2,x_3)$	
x_2	y_2		$y(x_1,x_2,x_3)$		$y(x_0,x_1,x_2,x_3,x_4)$
		$y(x_2,x_3)$		$y(x_1,x_2,x_3,x_4)$	
x_3	y_3		$y(x_2,x_3,x_4)$		
		$y(x_3,x_4)$			
x_4	y_4				

表示定理(representation theorem)

$$y(x_0,x_1,\cdots,x_n)=\sum_{i=0}^{n}\frac{y_1}{F_i'(x_i)}$$

表明,每一均差能被表示为 y_k 值的一个组合.这应与第 3 章中对应的定理相比较.

均差的对称性质是说,倘若 y_k 值以自变量 x_k 相同的方式排列,则这种差分是 x_k 的所有排列下的不变量.这个非常有用的结果是表示定理的一个简单推论.

均差与导数由

$$y(x,x_0,\cdots,x_n)=\frac{y^{(n+1)}(\xi)}{(n+1)!}$$

建立起联系.

在自变量等距的情况下,均差简化为常有限差分(ordinary finite difference),特别,

$$y(x_0,x_1,\cdots,x_n)=\frac{\Delta^n y_0}{n!h^n}.$$

据此,可得常有限差分的一个有用性质,即

$$\Delta^n y_0=y^{(n)}(\xi)h^n.$$

对于一个具有有界导数的函数 $y(x)$ 来说,所有的 $y^n(x)$ 有一个与 n 无关的界,因此,对于小的 h,当 n 无限增大时,有

$$\lim\Delta^n y_0=0.$$

这推广了前面得到的适合于多项式的结果,并且解释了为什么在一个差分表中,经常发现较高阶的差分趋于零.

现在,可以借助于均差获取配置多项式.典型的结果是 **Newton 均差公式**:

$$p(x)=y_0+(x-x_0)y(x_0,x_1)+(x-x_0)(x-x_1)y(x_0,x_1,x_2)$$
$$+\cdots+(x-x_0)(x-x_1)\cdots(x-x_{n-1})y(x_0,\cdots,x_n),$$

其中,自变量 x_k 不需要等距.这推广了第 6 章的 Newtow 公式,且在等距的情况下还原到它.

由于我们仍在讨论同样的配置多项式,所以误差 $y(x)-p(x)$ 仍由前面得到的公式

$$y(x)-p(x)=\frac{y^{(n+1)}(\xi)\pi(x)}{(n+1)!}$$

给出,其中,$y(x)$ 与 $p(x)$ 配置在自变量 $x_0,\cdots,x_n$ 上.利用均差,这个误差的另一种形式是

$$y(x)-p(x)=y(x,x_0,\cdots,x_n)(x-x_0)\cdots(x-x_n).$$

题　解

8.1 在数据点 $x = x_0, x_1, \cdots, x_n$ 上，Lagrange 乘子函数

$$L_i(x) = \frac{(x - x_0)(x - x_1)\cdots(x - x_{i-1})(x - x_{i+1})\cdots(x - x_n)}{(x_i - x_0)(x_i - x_1)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)}$$

取何值？

解 ☞ 首先注意到，对于 $k \neq i$，分子因式必使 $L_i(x_k) = 0$. 于是，分母因式必使 $L_i(x_i) = 1$.

8.2 验证：在自变量 $x_k, k = 0, \cdots, n$ 上，多项式 $p(x) = \sum_{i=0}^{n} L_i(x) y_i$ 取值为 y_k. 这是作为配置多项式的 Lagrange 公式.

证 ☞ 根据题 8.1，$p(x_k) = \sum_{i=0}^{n} L_i(x_k) y_i = L_k(x_k) y_k = y_k$，从而 Lagrange 公式的确给出了配置多项式.

8.3 用 $\pi(x)$ 定义乘积 $\pi(x) = \prod_{i=0}^{n}(x - x_i)$，试证

$$L_k(x) = \frac{\pi(x)}{(x - x_k)\pi'(x_k)}.$$

证 ☞ 由于 $\pi(x)$ 是 $n+1$ 个因子的乘积，通常的微分过程使 $\pi'(x)$ 产生 $n+1$ 项和，而每一项中有一个因子被求导. 如果我们定义

$$F_k(x) = \prod_{i \neq k}(x - x_i),$$

它除了比 $\pi(x)$ 缺少一个 $x - x_k$ 因子之外，其余相同，则

$$\pi'(x) = F_0(x) + \cdots + F_n(x).$$

于是，当 $x = x_k$ 时，由于 $F_k(x_k)$ 是惟一不包含 $x - x_k$ 的项，所以除了 $F_k(x_k)$，其余所有的项均为零. 从而

$$\pi'(x_k) = F_k(x_k) = (x_k - x_0)\cdots(x_k - x_{k-1})(x_k - x_{k+1})\cdots(x_k - x_n),$$

且

$$\frac{\pi(x)}{(x - x_k)\pi'(x_k)} = \frac{F_k(x)}{\pi'(x_k)} = \frac{F_k(x)}{F_k(x_k)} = L_k(x).$$

8.4 试证：行列式方程

$$\begin{vmatrix} p(x) & 1 & x & x^2 & \cdots & x^n \\ y_0 & 1 & x_0 & x_0^2 & \cdots & x_0^n \\ y_1 & 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y_n & 1 & x_n & x_n^2 & \cdots & x_n^n \end{vmatrix} = 0$$

也给出了配置多项式 $p(x)$.

证 ☞ 利用第一行元素的子式展开该行列式，将显然产生一个 n 次多项式. 用 $x = x_k$ 以及 $p(x) = y_k$ 代换使得两行元素相同从而行列式为零. 于是 $p(x_k) = y_k$，并且这个多项式是配置多项式. 值得注意的是，由于对大规模的行列式求值困难，这一结果用途不大.

8.5 找出取下面指定值的 3 次多项式.

x_k	0	1	2	4
y_k	1	1	2	5

解 ☞ 该多项式能直接被写为

$$p(x)=\frac{(x-1)(x-2)(x-4)}{(0-1)(0-2)(0-4)}\cdot 1+\frac{x(x-2)(x-4)}{1(1-2)(1-4)}\cdot 1$$
$$+\frac{x(x-1)(x-4)}{2(2-1)(2-4)}\cdot 2+\frac{x(x-1)(x-2)}{4(4-1)(4-2)}\cdot 5,$$

它能被整理为

$$p(x)=\frac{1}{12}(-x^3+9x^2-8x+12).$$

8.6 对于表 8.1 中的 y_k 值,计算直至 3 阶的均差.

表 8.1

x_k	y_k			
0	1			
		0		
1	1		$\frac{1}{2}$	
		1		$-\frac{1}{12}$
2	2		$\frac{1}{6}$	
		$\frac{3}{2}$		
4	5			

解 ☞ 该差分列在表的后三列中.例如

$$y(2,4)=\frac{5-2}{4-2}=\frac{3}{2},\qquad y(1,2,4)=\frac{\frac{3}{2}-1}{4-1}=\frac{1}{6},$$

$$y(0,1,2)=\frac{1-0}{2-0}=\frac{1}{2},\qquad y(0,1,2,4)=\frac{\frac{1}{6}-\frac{1}{2}}{4-0}=-\frac{1}{12}.$$

8.7 证明:$y(x_0,x_1)=y(x_1,x_0)$.这被称为一阶均差的**对称性**(symmetry).

证 ☞ 根据定义,这是显然的.但也可从

$$y(x_0,x_1)=\frac{y_0}{x_0-x_1}+\frac{y_1}{x_1-x_0}$$

这一事实看出.因为在这里,交换 x_0,x_1 以及 y_0,y_1,只需调换右端两项的顺序.这个过程可立即用于更高阶的差分.

8.8 证明 $y(x_0,x_1,x_2)$是对称的.

证 ☞ 将此差分改写为

$$y(x_0,x_1,x_2)=\frac{y(x_1,x_2)-y(x_0,x_1)}{x_2-x_0}=\frac{1}{x_2-x_0}\left(\frac{y_2-y_1}{x_2-x_1}-\frac{y_1-y_0}{x_1-x_0}\right)$$
$$=\frac{y_0}{(x_0-x_1)(x_0-x_2)}+\frac{y_1}{(x_1-x_0)(x_1-x_2)}+\frac{y_2}{(x_2-x_0)(x_2-x_1)},$$

交换任意两个自变量 x_j 与 x_k 及其对应的 y 值,即交换右端 y_j 与 y_k 项使总体结果不变.由于自变量 x_k 的任何置换会受相继的数对的交换影响,故在(x_k 与 y_k 双方的)所有置换下,均差是不变量.

8.9 证明:对任一正整数 n,

$$y(x_0,x_1,\cdots,x_n)=\sum_{i=0}^{n}\frac{y_i}{F_i^n(x_i)},$$

其中,

$$F_i^n(x_i)=(x_i-x_0)(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n).$$

这推广了前面两题的结果.

证 ☞ 证明采用归纳法 对于 $n=1,2$,我们已有此结果.假设对于 $n=k$,结果成立,则由定义,

$$y(x_0,x_1,\cdots,x_{k+1}) = \frac{y(x_1,\cdots,x_{k+1}) - y(x_0,\cdots,x_k)}{x_{k+1}-x_0}.$$

由于我们已假定对于 k 阶差分结果成立,故对于 $i=1,2,\cdots,k$,右端 y_k 的系数将是

$$\frac{1}{x_{k+1}-x_0}\left[\frac{1}{(x_i-x_1)\cdots(x_i-x_{k+1})} - \frac{1}{(x_i-x_0)\cdots(x_i-x_k)}\right],$$

这里,已知在分母乘积中不含因子 (x_i-x_i),而该系数简化为

$$\frac{1}{(x_i-x_0)\cdots(x_i-x_{k+1})} = \frac{1}{F_i^{k+1}(x_i)}$$

正如所求.当 $i=0$ 或 $i=k+1$ 时,y_i 的系数合二为一,然而在两种情况下容易看到当 $n=k+1$ 时,什么是定理中所需要的,这就是

$$\frac{1}{(x_0-x_1)\cdots(x_0-x_{k+1})} \quad \frac{1}{(x_{k+1}-x_0)\cdots(x_{k+1}-x_k)}$$

这就完成了归纳法并证明了该定理.

8.10 证明 n 阶均差是对称的.

证 ☞ 根据上述题立有如下结果:若交换任何一对自变量,比如 x_j 与 x_k,则右端含 y_j 与 y_k 的项交换而其余的均不变.

8.11 求 $y(x)=x^2$ 与 $y(x)=x^3$ 的前几个差分.

解 ☞ 首先取 $y(x)=x^2$,于是

$$y(x_0,x_1) = \frac{x_1^2-x_0^2}{x_1-x_0} = x_1+x_0,\ y(x_0,x_1,x_2) = \frac{(x_2+x_1)-(x_1+x_0)}{x_2-x_0} = 1.$$

显然,更高阶的差分将是零.现在取 $y(x)=x^3$

$$y(x_0,x_1) = \frac{x_1^3-x_0^3}{x_1-x_0} = x_1^2+x_1x_0+x_0^2,$$

$$y(x_0,x_1,x_2) = \frac{(x_2^2+x_2x_1+x_1^2)-(x_1^2+x_1x_0+x_0^2)}{x_2-x_0} = x_0+x_1+x_2,$$

$$y(x_0,x_1,x_2,x_3) = \frac{(x_1+x_2+x_3)-(x_0+x_1+x_2)}{x_3-x_0} = 1.$$

显然,更高阶的差分也将是零.注意,在这两种情况下,所有的差分均是对称的多项式.

8.12 证明,若 $k\leqslant n$,则一个 n 次多项式的 k 阶均差是一个 $n-k$ 次多项式;而若 $k>n$,则为零.

证 ☞ 记该多项式为 $p(x)$.一个典型的均差是

$$p(x_0,x_1) = \frac{p(x_1)-p(x_0)}{x_1-x_0}.$$

设想 x_0 固定而 x_1 是自变量,则这个公式的各部分能被看作 x_1 的函数.特别,分子是一个 x_1 的 n 次多项式,该多项式在 $x_1=x_0$ 时为零.根据因式定理,分子含 x_1-x_0 因子,因此其商 $p(x_0,x_1)$是 x_1 的一个 $n-1$ 次多项式.由 $p(x_0,x_1)$的对称性,它因此也是 x_0 的一个 $n-1$ 次多项式.重复相同的讨论,一个典型的2阶差分是

$$p(x_0,x_1,x_2) = \frac{p(x_1,x_2)-p(x_0,x_1)}{x_2-x_1}.$$

设想 x_0 与 x_1 固定而 x_2 是自变量,则分子是 x_2 的一个 $n-1$ 次多项式,该多项式在 $x_2=x_0$ 时为零.根据因式定理,$p(x_0,x_1,x_2)$因此是 x_2 的一个 $n-2$ 次多项式.由 $p(x_0,x_1,x_2)$的对称性,它也是 x_0 或 x_1 的 $n-2$ 次多项式.依次类推,能得到所需的结果.这需要用归纳法,但这是容易的,故略去细节.

8.13 证明 Newton 均差公式

$$p(x) = y_0 + (x-x_0)y(x_0,x_1) + (x-x_0)(x-x_1)y(x_0,x_1,x_2) + \cdots + (x-x_0)(x-x_1)\cdots(x-x_{n-1})y(x_0,\cdots,x_n)$$

表示配置多项式.即,对于 $k=0,\cdots,n$,它取值为 $p(x_k)=y_k$.

证 ☞ $p(x_0)=y_0$.是显然的事实.其次,由均差的定义,并利用对称性,

$$y_k=y_0+(x_k-x_0)y(x_0,x_k),$$
$$y(x_0,x_k)=y(x_0,x_1)+(x_k-x_1)y(x_0,x_1,x_k),$$
$$y(x_0,x_1,x_k)=y(x_0,x_1,x_2)+(x_k-x_2)y(x_0,x_1,x_2,x_k)$$
$$\cdots$$
$$y(x_0,\cdots,x_{n-2},x_k)=y(x_0,\cdots,x_{n-1})+(x_k-x_{n-1})y(x_0,\cdots,x_{n-1},x_k).$$

例如,由

$$y(x_0,x_1,x_k)=y(x_1,x_0,x_k)=\frac{y(x_0,x_k)-y(x_1,x_0)}{x_k-x_1}$$

得到第二行.对于 $k=1$,首先证明了 $p(x_1)=y_1$.将第二行代入第一行得到

$$y_k=y_0+(x_k-x_0)y(x_0,x_1)+(x_k-x_0)(x_k-x_1)y(x_0,x_1,x_k).$$

当 $k=2$ 时,这证明了 $p(x_2)=y_2$.对于每个 x_k,轮流逐次代入验证,直至最后,我们得到

$$y_n=y_0+(x_n-x_0)y(x_0,x_1)+(x_n-x_0)(x_n-x_1)y(x_0,x_1,x_2)$$
$$+\cdots+(x_n-x_0)(x_n-x_1)\cdots(x_n-x_{n-1})y(x_0,\cdots,x_{n-1},x_n),$$

这证明了 $p(x_n)=y_n$.

由于这个 Newton 公式与 Lagrange 公式表示相同的多项式,故两者中的每一个恰为另一个的重排.

8.14 找出在表 8.1 中取值的 3 次多项式.

解 ☞ 利用 Newton 公式,它包含表 8.1 的上端对角线上的差分,

$$p(x)=1+(x-0)0+(x-0)(x-1)\frac{1}{2}+(x-0)(x-1)(x-2)\left(-\frac{1}{12}\right).$$

简化为 $p(x)=\frac{1}{12}(-x^3+9x^2-8x+12)$.这与 Lagrange 公式得到的结果相同.

补 充 题

8.15 用 Lagrange 公式求一个包含如下 x_k、y_k 数对的 3 次多项式,然后对 $x=2,3,5$,求该多项式的值.

x_k	0	1	4	6
y_k	1	−1	1	−1

8.16 用 Lagrange 公式求一个包含如下 x_k,y_k 数对的 4 次多项式,然后对 $x=3$,求该多项式的值.

x_k	0	1	2	4	5
y_k	0	16	48	88	0

8.17 在部分分式展开式(partial fraction expansion)

$$\frac{p(x)}{\pi(x)}=\sum_{i=0}^{n}\frac{a_i}{x-x_i}$$

中,通过确定系数 a_i 来推出 Lagrange 公式(式子两端乘以 $x-x_i$,并令 x 以 x_i 为极限逼近 x_i,记住,就配置而言,$p(x_i)=y_i$),其结果是 $a_i=\frac{y_i}{\pi'(x_i)}$.

8.18 应用题 8.17 将 $\frac{3x^2+x+1}{x^3-6x^2+11x-6}$ 表示为部分分式的和

$$\frac{a_0}{x-x_0}+\frac{a_1}{x-x_1}+\frac{a_2}{x-x_2}.$$

(提示:对若干 x_0,x_1,x_2,想像分母为 $\pi(x)$,然后找出对应的 y_0,y_1,y_2,这等于将 $p(x)$ 当作一个配置多项式.)

8.19 将 $\frac{x^2+6x+1}{(x^2-1)(x-4)(x-6)}$ 表示为一个部分分式的和.

8.20　试证

$$L_0(x)=1+\frac{x-x_0}{x_0-x_1}+\frac{(x-x_0)(x-x_1)}{(x_0-x_1)(x_0-x_2)}+\cdots+\frac{(x-x_0)\cdots(x-x_{n-1})}{(x_0-x_1)\cdots(x_0-x_n)}$$

对于其他的系数,能根据对称性写出类似的展开式.

8.21　对于自变量 x_0, $x_0+\varepsilon$, x_1,写出3点 Lagrange 公式,然后讨论当 $\varepsilon\to0$ 时的极限.试证

$$p(x)=\frac{(x_1-x)(x+x_1-2x_0)}{(x_1-x_0)^2}y(x_0)+\frac{(x-x_0)(x_1-x)}{(x_1-x_0)}y'(x_0)$$

$$+\frac{(x-x_0)^2}{(x_1-x_0)^2}y(x_1)+\frac{1}{6}(x-x_0)^2(x-x_1)y'''(\xi).$$

借助于 $y(x_0)$, $y'(x_0)$与 $y(x_1)$,它定出一个2次多项式.

8.22　继上题,对于自变量 x_0, $x_0+\varepsilon$, $x_1-\varepsilon$, x_1,从 Lagrange 公式开始,借助于 $y(x_0)$, $y'(x_0)$, $y(x_1)$与 $y'(x_1)$,表示一个3次多项式.

8.23　对于如下的 x_k, y_k 数对,计算直至3阶的均差.

x_k	0	1	4	6
y_k	1	−1	1	1

8.24　对于题8.23中的 x_k, y_k 数对,利用 Newton 公式找出一个3次配置多项式.将你的结果与 Lagrange 公式所获得的结果进行比较.

8.25　重新排列题8.23的数对如下:

x_k	4	1	6	0
y_k	1	−1	−1	1

仍计算3阶均差.它将是与上面结果相同的描述对称性的数.

8.26　对如下的 y_k 值,计算一个4阶均差.

x_k	0	1	2	4	5
y_k	0	16	48	88	0

8.27　对于题8.26中的数据,应用 Newton 公式找配置多项式.在 $x=3$ 处,该多项式取何值?

8.28　试证

$$y(x_0,x_1)=\frac{\begin{vmatrix}1&y_0\\1&y_1\end{vmatrix}}{\begin{vmatrix}1&x_0\\1&x_1\end{vmatrix}},\quad y(x_0,x_1,x_2)=\frac{\begin{vmatrix}1&x_0&y_0\\1&x_1&y_1\\1&x_2&y_2\end{vmatrix}}{\begin{vmatrix}1&x_0&x_0^2\\1&x_1&x_1^2\\1&x_2&x_2^2\end{vmatrix}}.$$

8.29　对于 $y(x)=(x-x_0)(x-x_1)\cdots(x-x_n)=\pi(x)$,证明

$$y(x_0,x_1,\cdots,x_p)=0,\text{对 } p=0,1,\cdots,n;$$

$$y(x_0,x_1,\cdots,x_n,x)=1,\text{对所有 } x;$$

$$y(x_0,x_1,\cdots,x_n,x,z)=0,\text{对所有 } x,z.$$

8.30　对于 $k=-2,-1,0,1,2$,由验证 $p(x_k)=y_k$ 试证

$$p(x)=y_0+\frac{y(x_1,x_0)+y(x_0,x_{-1})}{2}(x-x_0)$$

$$+y(x_1,x_0,x_{-1})(x-x_0)\left(x-\frac{x_1+x_{-1}}{2}\right)$$

$$+\frac{y(x_2,x_1,x_0,x_{-1})+y(x_1,x_0,x_{-1},x_{-2})}{2}(x-x_1)(x-x_0)(x-x_{-1})$$

$$+y(x_2,x_1,x_0,x_{-1},x_{-2})(x-x_0)(x-x_1)(x-x_{-1})\left[x-\frac{x_2+x_{-2}}{2}\right]$$

是书写配置多项式的另一种方式.这是不等距的 Stirling 公式的一个推广,它能被推广到高次的.Bessel 公式与其他的公式也能被推广.

8.31 试证,对于等距的从而有 $x_{k+1}-x_k=h$ 的广变量,我们有

$$y(x_0,x_1,\cdots,x_n)=\frac{\Delta^n y_0}{n!h^n}.$$

8.32 带有两个或两个以上相等自变量的均差能由取极限的过程定义.例如,$y(x_0,x_0)$能定义为 $x\to x_0$ 时的极限 $\lim y(x,x_0)$,这可推出

$$y(x_0,x_0)=\lim\frac{y(x)-y_0}{x-x_0}=y'(x_0).$$

当 $y(x)=x^2$ 时,通过证明此时 $y(x,x_0)=x+x_0$ 从而 $\lim y(x,x_0)=y'(x_0)=2x_0$ 来直接验证它.当 $y(x)=x^3$ 时,由首先证明此时 $y(x,x_0)=x^2+xx_0+x_0^2$ 来同样直接地验证它.

8.33 在 2 阶均差

$$y(x_0,x,x_2)=\frac{y(x,x_2)-y(x_0,x_2)}{x-x_0}$$

中,右端可视为以 x_2 为一个常数的形式$\dfrac{f(x)-f(x_0)}{x-x_0}$.若 $\lim x=x_0$,我们定义

$$y(x_0,x_0,x_2)=\lim y(x_0,x,x_2).$$

由它推出

$$y(x_0,x_0,x_2)=y'(x,x_2)\big|_{x=x_0}.$$

当 $y(x)=x^3$ 时,由首先证明在这种情况下,当 $y(x,x_2)=x^2+xx_2+x_2^2$ 时

$$y(x_0,x,x_2)=x+x_0+x_2$$

而直接验证之.

第九章 样　　条

为了在整个区间上表示一个已知函数，使之达到所要求的精度，我们可以将若干个每个都是低次的多项式段联接在一起而不是用单个的可能是高次的多项式. 自然，古典的例子是以一组直线段，每一段都去拟合一个子区间上所给定的数据. 这种逼近是连续的，但在区间端点上有间断的一阶导数，成为角点(见图 9.1). 这就是用表值进行初等插值以及数值积分梯形法则的基础. 隐含的假设是在数据点之间已知函数几乎是线性的，如果数据点足够近的话，这可能是合理的.

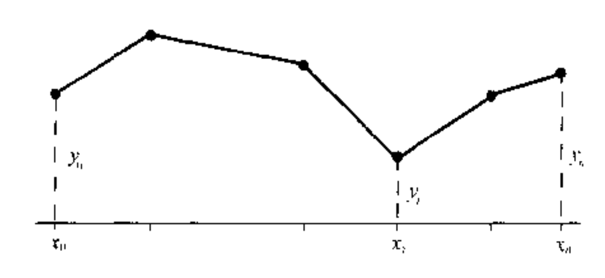


图 9.1　一个原始样条

在第 14 章中我们将要把抛物线段(二次多项式)拟合到一起来产生数值积分的辛甫生(Simpson)法则. 还将给出其他使用次数略高的多项式的例子. 在所有这些情况下，在线段联接处均有角点.

现在我们考虑一种方法，在这种方法中三次线段用一种方式拼合在一起，使角点圆化、在逼近时一阶和二阶导数均为连续的. 高次多项式有振动的特性. 一个 n 次的就最多可能有 $n-1$ 个转折点. 当以这样的一个多项式精确地表示一个已知函数时，它通常上下振动地经过该函数. 它有不想看到的边端效应，只有一阶导数可以逼近. 现在要导出的样条逼近就避免了这种振动，因为它是由低次线段所组成的. 样条这个词是来自于同名的绘图工具——绘制曲线时用的可任意弯曲的板条.

将已知区间 $(a,b)=I$ 用点 $x_0=a, x_1, x_2, \cdots, x_n=b$ 分成 n 个子区间，在每个子区间上以一个三次线段去拟合，在指定点 x_i 上取 y_i 值，在邻近区间的联接处有相同的一阶和二阶导数值. 点 x_1 直到 x_{n-1} 称为样条的节点或是结点(见图 9.2). 这些样条线段展开的细节将在题解中给出并且提供一些例子.

题　　解

9.1　一个三次(立方)多项式有四个系数，一般的表达式为

$$p(x)=c_0+c_1x+c_2x^2+c_3x^3,$$

通常如图 9.2 所示，n 个三次线段合在一起将含有 $4n$ 个系数. 它与施加在样条上的条件数目相比是什么情况呢?

解　☞ 值得注意的是通常我们期望 $4n$ 个系数可以用 $4n$ 个条件完全确定. 这里在每个结点 x_1 到 x_{n-1} 处要满足 4 个条件，即两侧的线段必须通过同一点而且一阶与二阶导数也要相同. 这就产生了 $4n-4$ 个条件. 在端点处只要求函数值相符，又增加了两个条件，合起来总共 $4n-2$ 个条件. 因此，样条并不能从已知的指定值完全地定义. 保留着两个自由度. 有时这些条件用来使端点的二阶导数为零，这就引出了所谓的自然样条. 另种方法是我们也可以要求端点的导数值与给定的函数导数值相符，如果这

些导数值是已知的或是可以予以逼近的话.第三种可以采用的方法是减少在结点 x_1 及 x_{n-1} 处的指定值.

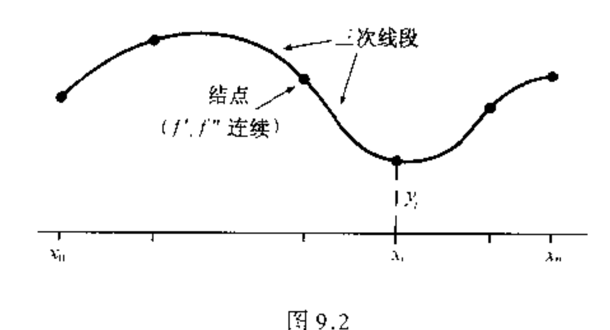


图 9.2

9.2 称图 9.2 中的子区间为 I_1 到 I_n,于是 $I_i=(x_{i-1},x_i)$.同时定义 $h_i=x_i-x_{i-1}$,注意子区间不必都是等长的.若 $S_i(x)$ 为 I_i 上的样条线段,证明对于常数 C_i 和 $i=1,\cdots,n$ 有

$$S_i''(x)=C_{i-1}\frac{x_i-x}{h_i}+C_i\frac{x-x_{i-1}}{h_i}.$$

证 ☞ 在 I_i 上样条段为三次的,所以它的一阶导数将是二次的而二阶导数为一次的.剩下要证实的是在每个结点 $x_k,k=1,\cdots,n-1$ 处的连续性.线段 S_k 在右端趋向这个结点而 S_{k+1} 在它的左端趋向这个结点.因此所要求的导数为

$$S_k''(x_k)=C_{k-1}\frac{x_k-x_k}{h_k}+C_k\frac{x_k-x_{k-1}}{h_k}$$

及

$$S_{k+1}''(x_k^+)=C_k\frac{x_{k+1}-x_k}{h_{k+1}}+C_{k+1}\frac{x_k-x_k}{h_{k+1}}.$$

二者均约化为 C_k.因此连续性得到了保证,而且我们发现常数 C_k 事实上就是样条二阶导数的共值.

9.3 将前题的结果积分二次就得到样条段,并且在样条段上加上通过相应结点的要求来确定积分常数.

解 ☞ 二次积分使得

$$S_i(x)=C_{i-1}\frac{(x_i-x)^3}{6h_i}+C_i\frac{(x-x_{i-1})^3}{6h_i}+c_i(x_i-x)+d_i(x-x_{i-1}),$$

其中后两项为积分常数所引入的线性函数,为了在结点处的定位,我们必须有 $S_i(x_{i-1})=y_{i-1}$ 及 $S_i(x_i)=y_i$.这些条件确定了 c_i 及 d_i,并导出

$$S_i(x)=C_{i-1}\frac{(x_i-x)^3}{6h_i}+C_i\frac{(x-x_{i-1})^3}{6h_i}+\left(y_{i-1}-\frac{C_{i-1}h_i^2}{6}\right)\frac{x_i-x}{h_i}$$
$$+\left(y_i-\frac{C_ih_i^2}{6}\right)\frac{x-x_{i-1}}{h_i},$$

这可以通过将 x_{i-1} 及 x_i 代入得到证实.

9.4 剩下的是保证一阶导数的连续性.要做到这一点,微分上一题的结果并且如同题 9.2 中那样比较邻近值.

解 ☞ 微分

$$S_i'(x)=-C_{i-1}\frac{(x_i-x)^2}{2h_i}+C_i\frac{(x-x_{i-1})^2}{2h_i}+\frac{y_i-y_{i-1}}{h_i}-\frac{C_i-C_{i-1}}{6}h_i.$$

因此在结点 x_k 处所要求的导数是

$$S_k'(x_k)=\frac{h_k}{6}C_{k-1}+\frac{h_k}{3}C_k+\frac{y_k-y_{k-1}}{h_k}$$

及

$$S_{k+1}'(x_k^+)=-\frac{h_{k+1}}{3}C_k-\frac{h_{k+1}}{6}C_{k+1}+\frac{y_{k+1}-y_k}{h_{k+1}}.$$

由于它们必须相等，对于 $k=1,\cdots,n-1$ 我们有

$$\frac{h_k}{6}C_{k-1}+\frac{h_k+h_{k+1}}{3}C_k+\frac{h_{k+1}}{6}C_{k+1}=\frac{y_{k+1}-y_k}{h_{k+1}}-\frac{y_k-y_{k-1}}{h_k},$$

它是常数 C_0 到 C_n 的 $n-1$ 个方程的线性方程组. 正如先前考察到的那样，这个方程组是不确定的. 我们还缺少两个方程.

存在令人感兴趣的办法来为这个方程组增添两个附加方程，既保持选择的开放性，其矩阵又保有一般特性. 首先对 $i=1,\cdots,n-1$，令

$$\alpha_i=\frac{h_{i+1}}{h_i+h_{i+1}},$$

$$\beta_i=1-\alpha_i=\frac{h_i}{h_i+h_{i+1}},$$

$$d_i=\frac{6}{h_i+h_{i+1}}\left(\frac{y_{i+1}-y_i}{h_{i+1}}-\frac{y_i-y_{i-1}}{h_i}\right).$$

这个方程组可以改写成

$$\beta_iC_{i-1}+2C_i+\alpha_iC_{i+1}=d_i,\qquad i=1,\cdots,n-1.$$

现在取两个附加条件其形式为

$$2C_0+\alpha_0C_1=d_0,\qquad \beta_nC_{n-1}+2C_n=d_n,$$

其中 α_0,d_0,β_n 和 d_n 可由我们支配. 这样构成的方程组形为

$$\begin{bmatrix}2&\alpha_0&0&&&&\\\beta_1&2&\alpha_1&&&&\\0&\beta_2&2&&&&\\&&&\ddots&&&\\&&&&2&\alpha_{n-2}&0\\&&&&\beta_{n-1}&2&\alpha_{n-1}\\&&&&0&\beta_n&2\end{bmatrix}\begin{bmatrix}C_0\\C_1\\C_2\\\vdots\\C_{n-2}\\C_{n-1}\\C_n\end{bmatrix}=\begin{bmatrix}d_0\\d_1\\d_2\\\vdots\\d_{n-2}\\d_{n-1}\\d_n\end{bmatrix}$$

该方程组矩阵为三对角的，所有其他元素为零.

9.5 怎样利用上题中的线性方程组来获得一个自然样条呢？

解 选择 α_0,d_0,β_n 及 d_n 为零. 这样一来顶行及底行的方程就迫使 C_0 及 C_n 也为零. 这正是自然样条. 此时系统缩成为 $n-1$ 阶，用来确定余下的 C_1 到 C_{n-1}.

9.6 类似地，我们怎样安排才能使端点条件

$$S_1'(x_0)=y_0',\quad S_n'(x_n)=y_n'$$

得以满足呢？

解 从题 9.4 中借用适当的公式我们可得

$$S_1'(x_0^+)=-\frac{h_1}{3}C_0-\frac{h_1}{6}C_1+\frac{y_1-y_0}{h_1}=y_0'$$

与

$$S_n'(x_n^-)=\frac{h_n}{6}C_{n-1}+\frac{h_n}{3}C_n+\frac{y_n-y_{n-1}}{h_n}=y_n',$$

容易将它们转变成

$$2C_0+C_1=\frac{6}{h_1}\left(\frac{y_1-y_0}{h_1}-y_0'\right)$$

与

$$C_{n-1}+2C_n=\frac{6}{h_n}\left(y_n'-\frac{y_n-y_{n-1}}{h_n}\right).$$

现在将它们与线性方程组中的第一个和最后一个方程(即 $2C_0+\alpha_0C_1=d_0$ 和 $\beta_nC_{n-1}+2C_n=d_n$)进行比较，得知可选择

$$\alpha_0=1=\beta_n,\quad d_0=\frac{6}{h_1}\left(\frac{y_1-y_0}{h_1}-y_0'\right),$$

$$d_n = \frac{6}{h_n}\left(y'_n - \frac{y_n - y_{n-1}}{h_n}\right).$$

事实上,这将提供所需的端点值.

9.7 在区间$(0,\pi)$上以三次样条段拟合函数 $f(x)=\sin x$. 只用二个内点 $\pi/3$ 及 $2\pi/3$.

解 ☞ 相应的数据组为

x_i	0	$\pi/3$	$2\pi/3$	π
y_i	0	$\sqrt{3}/2$	$\sqrt{3}/2$	0

取 $i=0,\cdots,3$ 及所有 $h_i=\pi/3$ 要寻求三个三次段. 由相等的 h_i 值立得 α_1、α_2、β_1 及 β_2 都等于 1/2. 于是

$$d_1 = \frac{3}{h}\left(\frac{0}{h} - \frac{\sqrt{3}/2}{h}\right) = -\frac{27\sqrt{3}}{2\pi^2},$$

并且 d_2 有相同的值. 这带给我们的是方程组

$$\frac{1}{2}C_0 + 2C_1 + \frac{1}{2}C_2 = \frac{-27\sqrt{3}}{2\pi^2},$$

$$\frac{1}{2}C_1 + 2C_2 + \frac{1}{2}C_3 = \frac{-27\sqrt{3}}{2\pi^2}.$$

及端点条件. 这里自然样条肯定适合,因为正弦函数在端点处的二阶导数恰为零. 所以我们令 C_0 及 C_3 为零. 于是余下的方程组立刻得出 $C_1=C_2=-27\sqrt{3}/5\pi^2$,代入题 9.3 的公式中终将产生样条段,将它简化后得

$$S_1(x) = \left(\frac{-27\sqrt{3}}{10\pi^3}\right)x^3 + \left(\frac{9\sqrt{3}}{5\pi}\right)x,$$

$$S_2(x) = \left(\frac{-27\sqrt{3}}{10\pi^3}\right)\left[\left(\frac{2\pi}{3} - x\right)^3 + \left(x - \frac{\pi}{3}\right)^3\right] + \frac{3\sqrt{3}}{5},$$

$$S_3(x) = \left(\frac{-27\sqrt{3}}{10\pi^3}\right)(\pi - x)^3 + \left(\frac{9\sqrt{3}}{5\pi}\right)(\pi - x).$$

题 9.19 将要求通过检查所有加给它们的条件来验证这些三次段. 此例的简明性使得精确值全都可以计算,还要注意中间一个"三次"段其实是二次的.

9.8 再一次用三次段去拟合正弦函数,这次要求端点的一阶导数与正弦函数的导数相等.

解 ☞ 新的端点条件为 $S'_1(0)=1$ 及 $S'_3(\pi)=-1$,从题 9.6 我们获得

$$\alpha_n = \beta_0 = 1 \quad d_0 = d_3 = \left(\frac{18}{\pi}\right)\left(\frac{3\sqrt{3}}{2\pi} - 1\right)$$

所以新的方程组为

$$2C_0 + C_1 = \left(\frac{18}{\pi}\right)\left(\frac{3\sqrt{3}}{2\pi} - 1\right),$$

$$\frac{1}{2}C_0 + 2C_1 + \frac{1}{2}C_2 = \frac{-27\sqrt{3}}{2\pi^2},$$

$$\frac{1}{2}C_1 + 2C_2 + \frac{1}{2}C_3 = \frac{-27\sqrt{3}}{2\pi^2},$$

$$C_2 + 2C_3 = \left(\frac{18}{\pi}\right)\left(\frac{3\sqrt{3}}{2\pi} - 1\right).$$

它的解为:

$$C_0 = C_3 = \frac{18\sqrt{3}}{\pi^2} - \frac{10}{\pi}, \quad C_1 = C_2 = \frac{2}{\pi} - \frac{9\sqrt{3}}{\pi^2}$$

将它代入题 9.3 的 $S_i(x)$的公式中去,我们再一次得到三次段. 如题 9.20 所要求的那样可以验证这些段满足所有加给它们的条件,在那里还可以发现 $S''(x)$的端点值不为零.

9.9 为获得样条逼近的一个完全确定的方程组,第三条途径是将我们的一些要求稍稍放宽. 例如,忽略 $S_1(x)$及 $S_n(x)$段,我们可以要求 $S_2(x)$和 $S_{n-1}(x)$满足端点的定位条件. 这还取消 S''' 在 x_1 及 x_{n-1}处的连续性要求,它们已不再是结点了. 证明所产生的问题将

拥有与确定系数所需的一样多的条件.

证 ☞ 现在有 $n-2$ 个三次段而不是 n 个,具有 $4n-8$ 个可用的系数.但是只有 $n-3$ 个结点而不是 $n-1$ 个.在每个结点上有 4 个要求,这就造成了 $4n-12$ 个需要满足的条件.由于在 x_0, x_1, x_{n-1} 及 x_n 处还需要定位,条件总数增至 $4n-8$ 个.

9.10 将题 9.2 中的展开修改成题 9.4 中的形式,来满足题 9.9 中所提出的要求.

解 ☞ 仔细阅读所提出的问题将表明有许多可以节省的地方.在题 9.4 中所列出的线性方程组里,当中 $n-3$ 个方程仍然有效,因为它们是关于结点 x_2 至 x_{n-2} 的,在那些点处一切依旧不变.这些已经提供了 $n-1$ 个系数 C_1 至 C_{n-1} 的 $n-3$ 个方程,另两个所需的方程将使 $S_2(x_0)=y_0$ 及 $S_{n-1}(x_n)=y_n$.回到题 9.3 中给出的公式,这些条件可以被实现.在进行某些代数操作之后它们可被简化成

$$2C_1+\alpha_1 C_2=d_1,\quad \beta_{n-1}C_{n-2}+2C_{n-1}=d_{n-1},$$

其中 $\alpha_1, \beta_{n-1}, d_1, d_{n-1}$ 定义如下:

$$\alpha_1=\frac{2(h_1h_2^2-h_1^3)}{(h_1+h_2)^3-(h_1+h_2)h_2^2},$$

$$\beta_{n-1}=\frac{2(h_{n-1}^2h_n-h_n^3)}{(h_{n-1}+h_n)^3-(h_{n-1}+h_n)h_{n-1}^2},$$

$$d_1=\frac{12h_2\left(y_0-\dfrac{h_1+h_2}{h_2}y_1+\dfrac{h_1}{h_2}y_2\right)}{(h_1+h_2)^3-(h_1+h_2)h_2^2},$$

$$d_{n-1}=\frac{12h_{n-1}\left(y_n-\dfrac{h_{n-1}+h_n}{h_{n-1}}y_{n-1}+\dfrac{h_n}{h_{n-1}}y_{n-2}\right)}{(h_{n-1}+h_n)^3-(h_{n-1}+h_n)h_{n-1}^2}.$$

于是,方程组的最后形式

$$\begin{bmatrix}2&\alpha_1&0&&&&\\ \beta_2&2&\alpha_2&&&&\\ 0&\beta_3&2&&&&\\ &&&\ddots&&&\\ &&&&2&\alpha_{n-3}&0\\ &&&&\beta_{n-2}&2&\alpha_{n-2}\\ &&&&0&\beta_{n-1}&2\end{bmatrix}\begin{bmatrix}C_1\\ C_2\\ C_3\\ \vdots\\ C_{n-3}\\ C_{n-2}\\ C_{n-1}\end{bmatrix}=\begin{bmatrix}d_1\\ d_2\\ d_3\\ \vdots\\ d_{n-3}\\ d_{n-2}\\ d_{n-1}\end{bmatrix}$$

仍然是三对角的,所有其他元素为零.

9.11 将适才展开的方法应用于区间 $(0,\pi)$ 上的函数 $f(x)=\sin x$,区间里取三个等距内点.

解 ☞ 有四个子区间,对内部的两个子区间寻求样条段.这里惟一结点是在 $x_2=\pi/2$ 处.这可以清楚地表明为什么我们不再继续前面的只有三个子区间的例子.在那里是在完全没有结点的情况下以单个三次式来插值四个已知点.如今的数据组为

x_i	0	$\pi/4$	$\pi/2$	$3\pi/4$	π
y_i	0	$\sqrt{2}/2$	1	$\sqrt{2}/2$	0

所有的 $h_i=\pi/4$. α_i 及 β_i 的公式现在只用在结点 x_2 上得出 $\alpha_2=\beta_2=1/2$.我们还得到 $d_2=48(\sqrt{2}-2)/\pi^2$ 以及一个方程

$$\frac{1}{2}C_1+2C_2+\frac{1}{2}C_3=\frac{48(\sqrt{2}-2)}{\pi^2}.$$

借助更新的公式,$\alpha_1=0, \beta_3=0$ 及

$$d_1=d_3=\frac{32(1-\sqrt{2})}{\pi^2}$$

于是我们的线性方程组如下:

$$
\begin{aligned}
2C_1 \qquad\qquad\qquad &= d_1,\\
\frac{1}{2}C_1 + 2C_2 + \frac{1}{2}C_3 &= \left(\frac{3\sqrt{2}}{2}\right)d_2,\\
2C_3 &= d_3.
\end{aligned}
$$

解之，并再一次借助题 9.3，我们便得到两个样条段：

$$
S_2(x) = \frac{16(1-\sqrt{2})(\pi-2x)^3 + (4\sqrt{2}-7)(4x-\pi)^3}{12\pi^3} + \frac{(8\sqrt{2}-2)(2\pi-4x)+(19-4\sqrt{2})(4x-\pi)}{12\pi},
$$

$$
S_3(x) = \frac{16(1-\sqrt{2})(2x-\pi)^3 + (4\sqrt{2}-7)(3\pi-4x)^3}{12\pi^3} + \frac{(8\sqrt{2}-2)(4x-2\pi)+(19-4\sqrt{2})(3\pi-4x)}{12\pi}.
$$

稍加耐心便可证实 S_2 联结前三个点，S_3 则是联结后三个，而且 X_2 是它们的正常结点. 这正是所求的一切. 额外再加些诸如 $S_2'(0)=1$ 或 $S_2''(\pi/2)=-1$ 这样的条件可能是好的，然而没有必要再做下去. 如今的逼近可以达到 1.05 和 −1.09.

9.12 样条逼近的误差是什么？

解 ☞ 可以证明

$$
\max\left|f(x)-S(x)\right| \leqslant \frac{5}{384}\max\left|f^{(4)}(x)\right| H^4.
$$

此处 H 为 h_i 的最大者而极大值是在区间 I 上取的.

9.13 将题 9.12 的误差界用于题 9.7 的样条上.

解 ☞ $\sin x$ 的 4 阶导数显然其界为 1 且 $H=\pi/3$. 因此

$$
\max\left|\sin x - S(x)\right| \leqslant \frac{5}{384}\,\frac{\pi^4}{81} = 0.016
$$

9.14 样条对导数 $f'(x)$ 的逼近是怎样的？

解 ☞ 可以证明

$$
\max\left|f'(x)-S'(x)\right| \leqslant \frac{\max\left|f^{(4)}(x)\right| H^3}{24}.
$$

9.15 将题 9.14 的公式应用到题 9.12 的样条上.

解 ☞ 我们近似地得到 $H^3/24=0.05$. 一般说来，样条对导数而言有十分好的逼近.

9.16 样条对 $f(x)$ 而言是全局逼近，其含义是什么？

解 ☞ 样条段不是彼此独立地确定的. 每段与其他段都联系在一起. 用来确定样条段的系数组 C_i 是由一个线性方程组来决定的. 用相反的方法，以头 4 点 x_0 到 x_3 来拟合一个三次多项式，然后对另一组 x_3 到 x_6 作拟合，如此等等遍及整个区间 I. 每个线段都独立于其他的去求得. 但是像样条那样在结点处的连续性几乎肯定会失去.

9.17 证明在所有具有二阶导数且在结点处满足 $f(x_i)=y_i$ 的函数 $f(x)$ 中，在 (a,b) 上的自然样条惟一地极小化

$$
\int_a^b f''(x)^2 \mathrm{d}x.
$$

证 ☞ 首先指出

$$
\int_a^b f''(x)^2\mathrm{d}x - \int_a^b S''(x)^2\mathrm{d}x = \int_a^b [f''(x)-S''(x)]^2\mathrm{d}x + 2\int_a^b S''(x)[f''(x)-S''(x)]\mathrm{d}x,
$$

其中 $S(x)$ 为三次样条. 在每个子区间上进行分部积分，最后一个积分就转变成

$$\int_{x_{i-1}}^{x_i} S_i''(x)[f'(x)-S_i'(x)]\mathrm{d}x$$

$$=S_i''(x)[f'(x)-S_i'(x)]\Big|_{x_{i-1}}^{x_i}-\int_{x_{i-1}}^{x_i}[f'(x)-S_i'(x)]S_i^{(3)}(x)\mathrm{d}x$$

$$=S_i''(x)[f'(x)-S_i'(x)]\Big|_{x_{i-1}}^{x_i}$$

$$-S_i^{(3)}(x)[f(x)-S_i(x)]\Big|_{x_{i-1}}^{x_i}$$

$$+\int_{x_{i-1}}^{x_i}[f(x)-S_i(x)]S_i^{(4)}(x)\mathrm{d}x.$$

由于 $f(x)$ 与 $S_i(x)$ 在结点处相等并且 $S_i^{(4)}(x)$ 为零.所以最后两项为零.将余下的一项对 $i=1,\cdots,n$ 求和,所有的内点相互抵消后剩下的是

$$S''(b)[f'(b)-S'(b)]-S''(a)[f'(a)-S'(a)].$$

由于 S 为自然样条它也为零.注意到若我们代之以假设在端点处 $f'=S'$,这剩下的项仍会消失.在任一种情况下将原始方程稍加重新整理便得

$$\int_a^b S''(x)^2\mathrm{d}x=\int_a^b f''(x)^2\mathrm{d}x-\int_a^b[f''(x)-S''(x)]^2\mathrm{d}x,$$

它确实使第一个积分比第二个小.

9.18 用一个三次样条来拟合下面的数据.

x_i	0	2	2.5	3	3.5	4	4.5	5	6
y_i	0	2.9	3.5	3.8	3.5	3.5	3.5	2.6	0

选择自然样条,题 9.4 中的 7 个方程的方程组提供关于 7 个内点的 C_i.它们的解,经过舍入到两位得

i	1	2	3	4	5	6	7
C_i	-0.23	-0.72	-4.08	2.65	0.69	-5.40	-0.70

解 ☞ 9 个数据点的分布情况及一个样条段如图 9.3 所示.回顾到 C_i 便是在数据点上的 2 阶导数值,C_0 与 C_8 为零.观察它们在整个区间的性态,特别是大的数值没有令人失望,可以消除人们的疑虑.

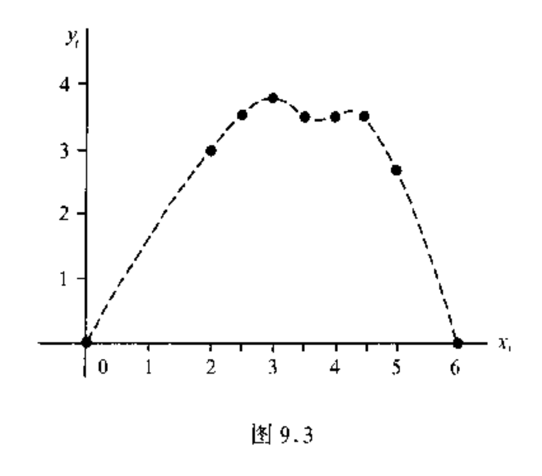


图 9.3

补 充 题

9.19 证明题 9.7 中的样条满足所有加给它的条件.

9.20　证明在题 9.8 中的第一个三次样条段为

$$S_1(x)=\frac{C_0}{2}\left(\frac{\pi}{3}-x\right)^3+\frac{C_1}{2}x^3-\frac{C_0\pi^2}{54}\left(1-\frac{3x}{\pi}\right)+\left(\frac{\sqrt{3}}{2}-\frac{C_1\pi^2}{54}\right)\frac{3x}{\pi},$$

并求出另外两个样条段.证明它们满足加给它们的要求.

9.21　证明在题 9.10 中给出的细节.

9.22　找出通过如下点的自然样条.

x_i	0	1	2	3	4
y_i	0	0	1	0	0

9.23　对前面的数据应用题 9.10 中的方法,在中间的两个区间上寻找两个样条段,惟一的一个结点在 $x=2$ 处,当然该样条也必须通过两个端点.

9.24　所有数据点均落在一根直线上的情况,它不能称为一个样条,然而它值得暂时给以注意.回想一下,常数 C_i 为 2 阶导数的值,在这种情况下必须都为零.我们的线性方程组应如何来处理这情况?

9.25　假如所有的数据点均落在一个抛物线上,我们的线性方程组将会是怎样的?

第十章　密切多项式

密切多项式不仅要求与给定的函数在若干指定的自变量处有相同的函数值(即配置的概念),而且要求在同样的自变量处(通常如此),它的直至某阶的导数也与给定函数的导数等值.因此,最简单的密切就是要求对于 $k=0,1,\cdots,n$ 有

$$p(x_k)=y(x_k),\qquad p'(x_k)=y'(x_k).$$

用几何语言来说,代表两个函数的曲线在这 $n+1$ 个点处彼此相切.高阶密切还要求 $p''(x_k)=y''(x_k)$,等等.这时相应的曲线就具有所谓的高阶接触.关于密切多项式的存在与惟一性,利用与那些用于简单定位多项式的相似的方法可以得到证明.

例如,Hermite 公式表示一个具有一阶密切的不超过 $2n+1$ 次的多项式.其形式为

$$p(x)=\sum_{i=0}^{n}U_i(x)y_i+\sum_{i=0}^{n}V_i(x)y_i'.$$

其中 y_i 与 y_i' 为给定函数和它的导数在 x_i 点处的值.函数 $U_i(x)$ 与 $V_i(x)$ 是多项式,它们与在早些时候所描述的 Lagrange 乘子具有相类似的性质.事实上

$$U_i(x)=[1-2L_i'(x_i)(x-x_i)][L_i(x)]^2,$$
$$V_i(x)=(x-x_i)[L_i(x)]^2.$$

Hermite 公式的误差可同样表示成位置误差的形式,不过带有高阶导数,这表明密切多项式可获得更高的精度.其误差为

$$y(x)-p(x)=\frac{y^{(2n+2)}(\xi)}{(2n+2)!}[\pi(x)]^2.$$

一种未定系数的方法可以用来获得具有高阶密切的多项式.例如,取 $p(x)$ 为标准形式

$$p(x)=c_0+c_1x+c_2x^2+\cdots+c_{3n+2}x^{3n+2},$$

并在点 $x_0,\cdots,x_n$ 处要求 $p(x_k)=y_k$, $p'(x_k)=y_k'$, $p''(x_k)=y_k''$,由此导出关于 $3n+3$ 个系数 c_i 的 $3n+3$ 个方程.无疑对于大的 n 来说将是一个大方程组.较后一章中的方法可以用来解这样的一个方程组.在某种情况下特殊的措施可以有效地加以简化.

题　解

10.1 假定(a) $U_i(x)$ 及 $V_i(x)$ 为 $2n+1$ 次多项式.(b) $U_i(x_k)=\delta_{ik}$, $V_i(x_k)=0$.(c) $U_i'(x_k)=0$, $V_i'(x_k)=\delta_{ik}$.

其中
$$\delta_{ik}=\begin{cases}0, & \text{当 } i\neq k,\\ 1, & \text{当 } i=k.\end{cases}$$

证明 $p(x)=\sum_{i=0}^{n}U_i(x)y_i+\sum_{i=0}^{n}V_i(x)y_i'$ 是一个次数不超过 $2n+1$ 的多项式,且满足 $p(x_k)=y_k$, $p'(x_k)=y'_k$.

证 ☞ 关于次数的结论是显见的,因为给定次数的多项式的迭加组合具有相同或是更低的次数.将 $x=x_k$ 代入便得

$$p(x_k)=U_k(x_k)y_k+0=y_k.$$

类似地将 $x=x_k$ 代入 $p'(x)$,

$$p'(x_k)=V_k'(x_k)y_k'=y_k',$$

所有其他项为零.

10.2 已知 Lagrange 乘子 $L_i(x)$ 满足 $L_i(x_k)=\delta_{ik}$,证明

$$U_i(x)=[1-2L_i'(x_i)(x-x_i)][L_i(x)]^2,$$
$$V_i(x)=(x-x_i)[L_i(x)]^2$$

满足列在题 10.1 中的要求.

证 ☞ 由于 $L_i(x)$ 为 n 次的,它的平方为 $2n$ 次的,因而 $U_i(x)$ 与 $V_i(x)$ 均为 $2n+1$ 次的.对第二个要求而言,我们注意到,由于 $L_i(x_k)=0$,故当 $k\neq i$ 时有 $U_i(x_k)=V_i(x_k)=0$.同时将 $x=x_i$ 代入,有

$$U_i(x_i)=[L_i(x_i)]^2=1,\quad V_i(x_i)=0,$$

于是有 $U_i(x_k)=\delta_{ik}$ 与 $V_i(x_k)=0$.接下来计算导数

$$U_i'(x)=[1-2L_i'(x_i)(x-x_i)]2L_i'(x)L_i(x)-2L_i'(x_i)[L_i(x)]^2$$
$$V_i'(x)=(x-x_i)2L_i(x)L_i'(x)+[L_i(x)]^2$$

由于有 $L_i(x_k)$ 这个因子立得.当 $k\neq i$ 时 $U_i'(x_k)=0$ 与 $V_i'(x_k)=0$.当 $x=x_i$ 时,由于 $L_i(x_i)=1$ 得 $U_i'(x_i)=2L_i'(x_i)-2L_i'(x_i)=0$.最后,$V_i'(x_i)=[L_i(x_i)]^2=1$.因此 Hermite 公式为

$$p(x)=\sum_{i=0}^{n}[1-2L_i'(x_i)(x-x_i)][L_i(x)]^2y_i$$
$$+(x-x_i)[L_i(x)]^2y_i'.$$

10.3 在平行的铁路轨道之间的一个岔道是一个联接位置(0,0)及(4,2)的三次多项式,并与 $y=0$ 以及 $y=2$ 这两根直线相切.如图 10.1 所示.应用 Hermite 公式来产生这个多项式.

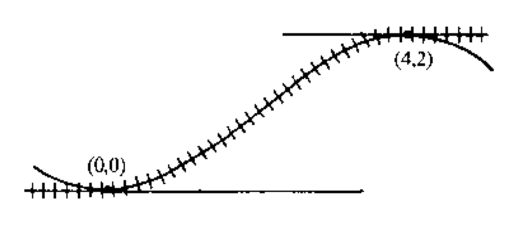


图 10.1

解 ☞ 这样的规定是要求一个三次多项式与下面的数据相匹配:

x_k	y_k	y'_k
0	0	0
4	2	0

当 $n=1$ 时,我们有

$$L_0(x)=\frac{x-x_1}{x_0-x_1},\quad L_1(x)=\frac{x-x_0}{x_1-x_0},$$
$$L_0'(x)=\frac{1}{x_0-x_1},\quad L_1'(x)=\frac{1}{x_1-x_0}.$$

代入 Hermite 公式(由于 $y_0=y_0'=y_1'=0$ 只有 y_1 项需要计算),

$$p(x)=\left(1-2\frac{x-4}{4-0}\right)\left(\frac{x-0}{4-0}\right)^2\cdot 2=\frac{1}{16}(6-x)x^2.$$

当然,这种转轨方式的意义在于它提供了一个光滑的过渡.它与两边的平行轨道都相切,也就没有突然的方向改变及角点.然而由于 $p''(0)$ 及 $p''(4)$ 不为零,曲率有间断.(另一种情况参看题 10.7)

10.4 获得 $y(x)$ 与它的多项式逼近 $p(x)$ 之差的公式.

解 ☞ 其推导十分类似于对较简单的配置多项式所作的那样.由于在点 $x_0,\cdots,x_n$ 上 $y(x)=p(x)$ 及 $y'(x)=p'(x)$,我们预期一个形如

$$y(x)-p(x)=C[\pi(x)]^2$$

的结果，其中 $\pi(x)=(x-x_0)\cdots(x-x_n)$ 如前，据此我们定义函数

$$F(x)=y(x)-p(x)-C[\pi(x)]^2,$$

它当 $k=0,\cdots,n$ 时有 $F(x_k)=F'(x_k)=0$，通过在 x_0 与 x_n 之间选择任一个新点 x_{n+1}，并使

$$C=\frac{y(x_{n+1})-p(x_{n+1})}{[\pi(x_{n+1})]^2}.$$

我们还使 $F(x_{n+1})=0$. 由于现在 $F(x)$ 至少有 $n+2$ 个零点，故 $F'(x)$ 在中间处会有 $n+1$ 个零点，且它在 $x_0,\cdots,x_n$ 处也有零点，合起来有 $2n+2$ 个零点，这就隐含了 $F''(x)$ 至少有 $2n+1$ 个零点，连续用 Rolle 定理，证明 $F^{(3)}(x)$ 至少有 $2n$ 个零点，$F^{(4)}(x)$ 有 $2n-1$ 个零点，依此类推直至 $F^{(2n+2)}(x)$ 可保证在 x_0 与 x_n 之间的区间中至少有一个零点，譬如说在 $x=\xi$ 处. 计算导数，我们得到

$$F^{(2n+2)}(\xi)=y^{(2n+2)}(\xi)-C(2n+2)!=0.$$

它可以就 C 解出，回代得

$$y(x_{n+1})-p(x_{n+1})=\frac{y^{(2n+2)}(\xi)}{(2n+2)!}[\pi(x_{n+1})]^2.$$

回顾 x_{n+1} 为 $x_0,\cdots,x_n$ 外的另一个点，并注意到这结果即使对 $x_0\cdots x_n$ 也成立(两边均为零)，故可以将 x_{n+1} 换写成简单一些的 x：

$$y(x)-p(x)=\frac{y^{(2n+2)}(\xi)}{(2n+2)!}[\pi(x)]^2.$$

10.5 证明能满足题 10.1 中规定条件的多项式是惟一的.

证 ☞ 假设有两个的话，由于他们在点 x_k 处 y_k 及 y'_k 的值必须相同，我们可以选择其中的一个为题 10.4 中的 $p(x)$ 而另一个为 $y(x)$. 换言之，我们可以把一个多项式看成另一个的逼近式，然而由于 $y(x)$ 现在是一个 $2n+1$ 次的多项式，由此得出 $y^{(2n+2)}(\xi)$ 为零. 因此，$y(x)$ 恒等于 $p(x)$，从而我们的两个多项式实质上就是一个.

10.6 怎样才能找到一个多项式与下面的数据相匹配？

$$\begin{matrix} x_0 & y_0 & y'_0 & y''_0 \\ x_1 & y_1 & y'_1 & y''_1 \end{matrix}$$

换言之，在二个点上多项式和它的头两阶导数取指定值.

解 ☞ 为了简单化可令 $x_0=0$. 假如这一点不成立，对点加以平移容易实现它. 令

$$p(x)=y_0+xy'_0+\frac{1}{2}x^2y''_0+Ax^3+Bx^4+Cx^5$$

其中 A,B 和 C 是要确定的. 在 $x=x_0=0$ 处指定值已经被满足. 在 $x=x_1$ 处它们要求

$$Ax_1^3+Bx_1^4+Cx_1^5=y_1-y_0-x_1y'_0-\frac{1}{2}x_1^2y''_0$$

$$3Ax_1^2+4Bx_1^3+5Cx_1^4=y'_1-y'_0-x_1y''_0$$

$$6Ax_1+12Bx_1^2+20Cx_1^3=y''_1-y''_0$$

这三个方程惟一地确定 A,B,C.

10.7 在两条平行的铁路轨道之间的一个转轨岔道将位置(0,0)和(4,2)联接在一起. 为了避免在方向和曲率二者上的不连续作下面的规定：

x_k	y_k	y'_k	y''_k
0	0	0	0
4	2	0	0

求一个多项式它满足这些规定值.

解 ☞ 应用题 10.6 中的过程

$$p(x)=Ax^3+Bx^4+Cx^5$$

二次项完全地消失. 在 $x_1=4$ 处我们得到

$$64A + 256B - 1024C = 2, \quad 48A + 256B + 1280C = 0,$$
$$24A + 192B - 1280C = 0.$$

由此得 $A = \frac{40}{128}, B = -\frac{15}{128}, C = \frac{3}{256}$, 代入得

$$p(x) = \frac{1}{256}(80x^3 - 30x^4 + 3x^5).$$

补 充 题

10.8 应用 Hermite 公式来求一个二次多项式，它满足下面这些规定值.

x_k	y_k	y'_k
0	0	0
1	1	1

这可以看成是在非平行轨道之间的转轨岔道

10.9 应用 Hermite 公式来求一个多项式，它满足下面这些规定值.

x_k	y_k	y'_k
0	0	0
1	1	0
2	0	0

10.10 应用题 10.6 中的方法求一个满足下面这些规定值的 5 次多项式.

x_k	y_k	y'_k	y''_k
0	0	0	0
1	1	1	0

这是一个较题 10.8 更光滑的转轨岔道.

10.11 求二个二次多项式，其一具有 $p_1(0) = p'_1(0) = 0$，另一个具有 $p_2(4) = 2, p'_2(4) = 0$，二者均通过(2, 1)，如图 10.2 所示，证明 $p'_1(2) = p'_2(2)$，使得一对抛物线弧也能作为平行轨道之间的转轨岔道，就如同题 10.3 中的三次曲线一样.

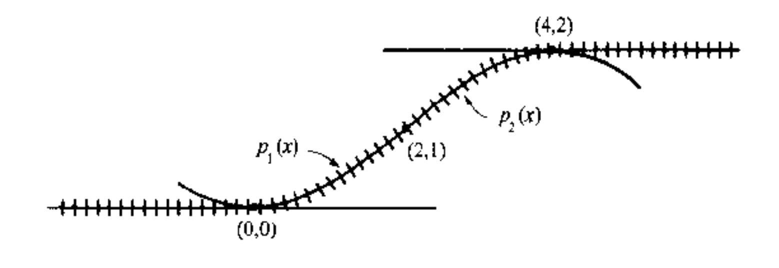


图 10.2

10.12 寻找二个 4 次多项式，一个具有 $p_1(0) = p'_1(0) = p''_1(0) = 0$，另一个具有 $p_2(4) = 2, p'_2(4) = p''_2(4) = 0$，二者均通过(2, 1)满足 $p''_1(2) = p''_2(2) = 0$. 这是另一种转轨岔道，对它来说如同题 10.7 的 5 次多项式那样，方向及曲率均没有间断，通过证明在四片轨道的衔接点(0, 0)，(2, 1)和(4, 2)的两侧一阶和二阶导数都相等来证明这一点.

10.13 从 2 点密切的 Hermite 公式来导出中点公式

$$p_{1/2} = \frac{1}{2}(y_0 + y_1) + \frac{1}{8}L(y'_0 - y'_1)$$

其中 $L = x_1 - x_0$.

10.14　证明题 10.13 中公式的误差为 $L^4y^{(4)}(\xi)/384$.

10.15　求一个 4 次多项式满足下面的条件：

x_k	y_k	y'_k
0	1	0
1	0	—
2	9	24

注意 y'_k 中的一个值是不用的.

10.16　求一个 4 次多项式满足这些条件：

x_k	y_k	y'_k	y''_k
0	1	−1	0
1	2	7	—

10.17　求一个 3 次多项式满足这些条件：

x_k	y_k	y'_k
0	1	−2
1	1	4

第十一章 Taylor 多项式

Taylor 多项式

在密切多项式中首推 Taylor 多项式.在单个点 x_0 处多项式及它的头 n 阶导数值被要求与给定函数 $y(x)$的有关量相匹配.那就是

$$p^{(i)}(x_0) = y^{(i)}(x_0) \qquad 当\ i = 0, 1, \cdots, n.$$

这样一个多项式的存在性与惟一性将被证明,它是分析学的经典结果.Taylor 公式的存在性问题可以通过将它展开成如下形式

$$p(x) = \sum_{i=0}^{n} \frac{y^{(i)}(x_0)}{i!}(x - x_0)^i$$

而得到肯定.

当 Taylor 公式作为对 $y(x)$的逼近时,其误差可以表示为积分形式

$$y(x) - p(x) = \frac{1}{n!}\int_{x_0}^{x} y^{(n+1)}(x_0)(x - x_0)^n \mathrm{d}x_0.$$

Lagrange 误差公式可以通过对该积分公式应用一个中值定理演绎出来.它就是

$$y(x) - p(x) = \frac{y^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1},$$

而且明显地类似于配置公式与密切公式的误差.

假如 $y(x)$的导数界与 n 无关,则任一个误差公式都可用来估计所需要的次数 n 以使误差$|y(x) - p(x)|$在自变量 x 的整个给定区间上减少到预先给定的允许量之下.

解析函数具有这样的性质,即当 n 趋向无穷时,上面的逼近误差对给定区间内的所有自变量 x 其极限为零.这类函数于是可以表示为 Taylor 级数

$$y(x) = \sum_{i=0}^{\infty} \frac{y^{(i)}(x_0)}{i!}(x - x_0)^i.$$

二项式级数为 Taylor 级数的一个特别重要的情况,当 $-1 < x < 1$ 时我们有

$$(1 + x)^p = \sum_{i=0}^{\infty} \binom{p}{i} x^i.$$

微分算子 D

微分算子 D 定义为

$$D = h\frac{\mathrm{d}}{\mathrm{d}x}.$$

于是指数算子可以定义如下:

$$e^{kD} = \sum_{i=0}^{\infty} \frac{k^i D^i}{i!}$$

并且 Taylor 级数可表述为算子形式

$$y(x_k) = e^{kD} y_0(x_0).$$

D 与 Δ 之间的关系可用

$$\Delta + 1 = e^D, \quad D = \Delta - \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 - \cdots,$$

二者中的任一种来表示,二者均包括"无穷级数"算子.

Euler 变换为无穷级数算子间的另一个有用关系,它可以用二项式级数写成

$$(1+E)^{-1}=\frac{1}{2}\left[1-\frac{1}{2}\Delta+\frac{1}{4}\Delta^2-\frac{1}{8}\Delta^3+\cdots\right].$$

Bernoulli 数 B_i 定义为

$$\frac{x}{e^x-1}=\sum_{i=0}^{\infty}\frac{1}{i!}B_ix^i,$$

着手将左侧展成它的 Taylor 级数，我们将得到 $B_0=1, B_1=-\frac{1}{2}, B_2=\frac{1}{6}$ 等等. 这些数出现在不同的算子方程中，例如，不定和算子 Δ^{-1}定义为

$$\Delta F_k=y_k,\quad F_k=\Delta^{-1}y_k.$$

与 D 的关系为

$$\Delta^{-1}=D^{-1}\sum_{i=0}^{\infty}\frac{1}{i!}B_iD^i,$$

其中 B_i 为 Bernoulli 数. 算子 D^{-1}为熟悉的不定积分算子.

从前面的关系式中可以推导出 Euler-Maclaurin 公式

$$\sum_{i=0}^{n-1}y_i=\int_0^n y_k\mathrm{d}k-\frac{1}{2}(y_n-y_0)+\frac{h}{12}(y_n'-y_0')+\cdots,$$

它常用来计算和数与积分.

使用 Taylor 级数可以把 D 的幂以中心差分算子 δ 来表示. 下面是一些例子

$$D=\mu\left(\delta-\frac{1^2}{3!}\delta^3+\frac{1^2\cdot 2^2}{5!}\delta^5-\frac{1^2\cdot 2^2\cdot 3^2}{7!}\delta^7+\cdots\right),$$

$$D^2=\delta^2-\frac{1}{12}\delta^4+\frac{1}{90}\delta^6-\frac{1}{560}\delta^8+\frac{1}{3150}\delta^{10}-\cdots.$$

题　解

11.1　找出一个次数为 n 或更低次的多项式，使它与它的前 n 次导数在自变量 x_0 处取值 $y_0, y_0^{(1)}, y_0^{(2)}, \cdots, y_0^{(n)}$

解　☞ 一个 n 次的多项式可以写成

$$p(x)=a_0+a_1(x-x_0)+\cdots+a_n(x-x_0)^n,$$

逐次微分的结果为

$$p^{(1)}(x)=a_1+2a_2(x-x_0)+\cdots+na_n(x-x_0)^{n-1},$$

$$p^{(2)}(x)=2a_2+3\cdot 2a_3(x-x_0)+\cdots+n(n-1)a_n(x-x_0)^{n-2},$$

$$\cdots$$

$$p^{(n)}(x)=n!a_n.$$

于是那些规定值要求

$$p(x_0)=a_0=y_0,\quad p^{(1)}(x_0)=a_1=y_0^{(1)},\quad p^{(2)}(x_0)=2a_2=y_0^{(2)},$$

$$\cdots p^{(n)}(x_0)=n!a_n=y_0^{(n)}.$$

解出 a_n 系数并回代得

$$p(x)=y_0+y_0^{(1)}(x-x_0)+\cdots+\frac{1}{n!}y_0^{(n)}(x-x_0)^n$$

$$=\sum_{i=0}^{n}\frac{1}{i!}y_0^{(i)}(x-x_0)^i.$$

11.2　找出一个 n 次多项式 $p(x)$，要求于 $x_0=0$ 处 $p(x)$与 e^x 直至 n 阶导数的值均相同.

解　☞ 由于对 e^x 而言其所有阶的导数仍为 e^x，故

$$y_0=y_0^{(1)}=y_0^{(2)}=\cdots=y_0^{(n)}=1$$

于是 Taylor 多项式可以写成

$$p(x) = \sum_{i=0}^{n} \frac{1}{i!}x^n = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots + \frac{1}{n!}x^n.$$

11.3 考虑第二个函数 $y(x)$ 也具有题 11.1 中的规定值.我们把 $p(x)$ 看成是逼近 $y(x)$ 的一个多项式.假定 $y^{(n+1)}(x)$ 在 x_0 与 x 之间连续求得一个以积分形式表示的差 $y(x) - p(x)$ 的公式.

解 ☞ 这里用一个不同的过程来导出关于配置多项式和密切多项式的误差估计是方便的.首先临时称这个差为 R:

$$R = y(x) - p(x),$$

或者将它全部的细节写出

$$R(x,x_0) = y(x) - y(x_0) - y'(x_0)(x - x_0) - \frac{1}{2}y''(x_0)(x - x_0)^2$$
$$- \cdots - \frac{1}{n!}y^{(n)}(x_0)(x - x_0)^n,$$

这实际上定义 R 作为 x 和 x_0 的函数.保持 x 固定,计算 R 对 x_0 的导数,我们得到

$$\begin{aligned} R'(x,x_0) = & -y'(x_0) + y'(x_0) - y''(x_0)(x - x_0) \\ & + y''(x_0)(x - x_0) - \frac{1}{2}y^{(3)}(x_0)(x - x_0)^2 + \cdots \\ & - \frac{1}{n!}y^{(n+1)}(x_0)(x - x_0)^n \\ = & -\frac{1}{n!}y^{(n+1)}(x_0)(x - x_0)^n \end{aligned}$$

由于在每项乘积中,对第二个因子进行微分的结果会与在前一项乘积中对第一个因子微分的结果相消,因此只有最后一项保留下来.已对 x_0 作了微分现在反过来对 x_0 进行积分以重新获得 R

$$R(x,x_0) = -\frac{1}{n!}\int_x^{x_0} y^{(n+1)}(u)(x - u)^n \mathrm{d}u + 常数$$

由 R 的最初定义,有 $R(x_0,x_0) = 0$,从而积分常数为零,将上下限互换得

$$R(x,x_0) = \frac{1}{n!}\int_{x_0}^{x} y^{(n+1)}(u)(x - u)^n \mathrm{d}u,$$

这就是积分形式的误差公式.

11.4 从积分形式的误差来得到 Lagrange 形式的误差.

解 ☞ 这里我们用微积分学中的中值定理,该定理说假如 $f(x)$ 为连续的且 $w(x)$ 在区间 (a,b) 中符号不变,则

$$\int_a^b f(x)w(x)\mathrm{d}x = f(\xi)\int_a^b w(x)\mathrm{d}x.$$

这里 ξ 为 a 与 b 之间的一个值,选 $w(x) = (x - x_0)^n$,我们容易地得到

$$R(x,x_0) = \frac{1}{(n+1)!}y^{(n+1)}(\xi)(x - x_0)^{n+1},$$

其中 ξ 是介于 x_0 与 x 之间的量但不能说是未知的.这种误差的形式非常流行,因为它十分类似于 Taylor 多项式的项.除了将 x_0 换成 ξ,它也就是更高一次的 Taylor 多项式的项.

11.5 对函数 $y(x) = e^x$,取 $x_0 = 0$,估计对于 $-1 < x < 1$ 保证准确到 3 位小数的 Taylor 多项式次数以及准确到 6 位的次数.

解 ☞ 用误差的 Lagrange 公式

$$|e^x - p(x)| = |R| \leqslant \frac{e}{(n+1)!}$$

要准确到 3 位它必须不超过 0.0005,$n = 7$ 或更高些可以满足这个要求.因此多项式

$$p(x) = \sum_{i=0}^{7} \frac{1}{i!}x^i$$

是可以胜任的.类似地准确到 6 位 R 必须不超过 0.0000005,当 $n = 10$ 时满足要求.

11.6 算子 D 的定义为 $D = h\dfrac{\partial}{\partial x}$.将它的逐次幂作用于 $y(x)$ 上其结果为何?

解 ☞ 我们应得 $D^iy(x)=h^iy^{(i)}(x)$.

11.7 用算子符号表示 Taylor 多项式.

解 ☞ 令 $x-x_0=kh$ 这是我们早些时用过的符号现在简单地以 x 代替 x_k. 然后直接代入题 11.1 中的 Taylor 多项式便得

$$p(x)=\sum_{i=0}^{n}\frac{1}{i!}y_0^{(i)}(x-x_0)^i=\sum_{i=0}^{n}\frac{1}{i!}y_0^{(i)}k^ih^i$$
$$=\sum_{i=0}^{n}\frac{1}{i!}k^iD^iy(x_0).$$

改写该结果的一个常用的方式是

$$p(x)=\left(\sum_{i=0}^{n}\frac{1}{i!}k^iD^i\right)y(x_0),$$

或单用整变量 k 就成为

$$p_k=\left(\sum_{i=0}^{n}\frac{1}{i!}k^iD^i\right)y_0.$$

此处如通常的那样 $p(x_k)=p_k$.

11.8 一个函数 $y(x)$ 在区间 $|x-x_0|\leqslant r$ 上称为解析的,若当 $n\to\infty$ 时

$$\lim R(x,x_0)=0$$

对区间内的所有自变量 x 成立. 于是习惯地将 $y(x)$ 写成一个无穷级数,称之为 Taylor 级数

$$y(x)=\lim p(x)=\sum_{i=0}^{\infty}\frac{1}{i!}y_0^{(i)}(x-x_0)^i,$$

将它表示为算子形式.

解 ☞ 就像在题 11.7 中那样做的,我们得到 $y(x_k)=\left(\sum_{i=0}^{\infty}\frac{1}{i!}k^iD^i\right)y_0$. 这是我们的第一个"无穷级数算子". 此类算子的算术运算不像早些时对简单算子所做的那样容易证明它的正确性.

11.9 算子 e^{kD} 由 $e^{kD}=\sum_{i=0}^{\infty}\frac{1}{i!}k^iD^i$ 所定义. 用这个算子来写 Taylor 级数.

解 ☞ 我们立得 $y(x_k)=e^{kD}y_0$.

11.10 证明 $e^D=E$.

证 ☞ 在题 11.9 中取 $k=1$ 并据 E 的定义,有 $y(x_1)=y_1=Ey_0=e^Dy_0$, 便得 $E=e^D$.

11.11 将 $y(x)=\ln(1+x)$ 展成 Taylor 级数,取 $x_0=0$.

解 ☞ 导数为 $y^{(i)}(x)=(-1)^{i+1}(i-1)!\ /(1+x)^i$, 于是 $y^{(i)}(0)=(-1)^{i+1}(i-1)!$. 由于 $y(0)=\ln 1=0$, 我们得到

$$y(x)=\ln(1+x)=\sum_{i=1}^{\infty}\frac{(-1)^{i+1}}{i}x^i=x-\frac{1}{2}x^2+\frac{1}{3}x^3-\frac{1}{4}x^4+\cdots.$$

用熟悉的比率检验法可证明级数在 $-1<x<1$ 中收敛. 然而,这并不证明该级数等于 $\ln(1+x)$. 为了证明这一点令 $p(x)$ 表示 n 次的 Taylor 多项式. 然后由误差的 Lagrange 公式得

$$|\ln(1+x)-p(x)|\leqslant\frac{1}{(n+1)!}\cdot\frac{n!}{(1+\xi)^{n+1}}\cdot x^{n+1}.$$

为简单起见只考虑区间 $0\leqslant x<1$. 该级数无疑多半是用于这个区间的. 于是误差可以通过以 0 代替 ξ 以 1 代替 x 得到 $|\ln(1+x)-p(x)|\leqslant 1/(n+1)$ 来进行估计. 而它的极限确为零. 因此, $\lim p(x)=\ln(1+x)$, 这正是我们所要的.

11.12 估计以一个 Taylor 多项式在 $x_0=0$ 处来逼近 $y(x)=\ln(1+x)$ 所需的次数,使其在 $0<x<1$ 中保证有 3 位小数的精度.

解 ☞ 用 Lagrange 误差公式

$$|\ln(1+x)-p(x)|\leqslant\frac{1}{(n+1)!}\cdot\frac{n!}{(1+\xi)^n}\cdot x^{n+1}\leqslant\frac{1}{n+1},$$

3 位精确度要求这个界不超过 0.0005，要 $n=2000$ 或更高些才得到满足．一个 2000 次的多项式是必须的，这是一个慢收敛级数的例子．

11.13 以 Δ 算子来表示算子 D.

解 ☞ 从 $e^D=E$ 我们得到 $D=\ln E=\ln(1+\Delta)=\Delta-\frac{1}{2}\Delta^2+\frac{1}{3}\Delta^3-\frac{1}{4}\Delta^4+\cdots$.

该计算的有效性肯定还存有疑问，对它的任何应用必须仔细地检查．本题表明最后的级数算子会产生与算子 D 相同的结果．

11.14 将 $y(x)=(1+x)^p$ 表为 Taylor 级数．

解 ☞ 当 p 为正整数时，它是代数的二项式定理，当 p 为其他值时，它就是二项式级数．它的应用是广泛的，我们容易得到

$$y^{(i)}(x)=p(p-1)\cdots(p-i+1)(1+x)^{p-i}=p^{(i)}(1+x)^{p-i},$$

其中 $p^{(i)}$ 再次为阶乘多项式，选 $x_0=0$，

$$y^{(i)}(0)=p^{(i)},$$

将它代入 Taylor 级数

$$y(x)=\sum_{i=0}^{\infty}\frac{p^{(i)}}{i!}x^i=\sum_{i=0}^{\infty}\binom{p}{i}x^i,$$

其中 $\binom{p}{i}$ 为广义的二项式系数．这个级数在 $-1<x<1$ 中收敛于 $y(x)$ 是可以证实的．

11.15 用二项式级数导出 Euler 变换．

解 ☞ Euler 变换是交错级数 $S=a_0-a_1+a_2-a_3+\cdots$ 的一个拓广的重新排列，通过以 $p=-1$ 的二项式定理我们可以将 S 改写为

$$S=(1-E+E^2-E^3+\cdots)a_0=(1+E)^{-1}a_0.$$

算子 $(1+E)^{-1}$ 可以解释为 $1+E$ 的逆算子．下面是二项式定理的第二个应用．

$$S=(1+E)^{-1}a_0=(2+\Delta)^{-1}a_0=\frac{1}{2}\left(1+\frac{\Delta}{2}\right)^{-1}a_0$$

$$=\frac{1}{2}\left(1-\frac{\Delta}{2}+\frac{\Delta^2}{4}-\frac{\Delta^3}{8}+\cdots\right)a_0$$

$$=\frac{1}{2}\left(a_0-\frac{1}{2}\Delta a_0+\frac{1}{4}\Delta^2a_0-\frac{1}{8}\Delta^3a_0+\cdots\right).$$

我们推导这个公式曾是算子算术的多少有点乐观的应用．不总存在易于应用的判别准则来保证它的有效性．

11.16 下面级数中的那些数 B_i 定义为 Bernoulli 数：

$$y(x)=\frac{x}{e^x-1}=\sum_{i=0}^{\infty}\frac{1}{i!}B_ix^i.$$

求 $B_0,\cdots,B_{10}$.

解 ☞ Taylor 级数要求 $y^{(i)}(0)=B_i$，然而在这种情况下走另一条路要更容易些．乘以 e^x-1 并用 e^x 的 Taylor 级数，我们得到

$$x=\left(x+\frac{1}{2}x^2+\frac{1}{6}x^3+\cdots\right)\left(B_0+B_1x+\frac{1}{2}B_2x^2+\frac{1}{6}B_3x^3+\cdots\right).$$

现在比较 x 逐次幂的系数，得到

$$B_0=1,B_1=-\frac{1}{2},B_2=\frac{1}{6},B_3=0,B_4=-\frac{1}{30},B_5=0,$$

$$B_6=\frac{1}{42},B_7=0,B_8=-\frac{1}{30},B_9=0,B_{10}=\frac{5}{60}.$$

这个过程可以用显见的方式继续下去．

11.17 假设 $\Delta F_k=y_k$．那么其逆算子可以定义为 $F_k=\Delta^{-1}y_k$．逆算子是“不定的”由于对给定的 y_k 除了一个任意的常数外才是确定的．例如，在下面的表中 y_k 作为一阶差分被列出，证明可以选 F_0 为任意的，而其他的 F_k 就确定了．

F_k	F_0	·	·	·	·	·	·	·
y_k	y_0	y_1	y_2	y_3	y_4	·	·	·

证 ☞ 我们立刻得到

$$F_1 = F_0 + y_0,\quad F_2 = F_1 + y_1 = F_0 + y_0 + y_1,$$
$$F_3 = F_2 + y_2 = F_0 + y_0 + y_1 + y_2.$$

一般地有 $F_k = F_0 + \sum_{i=0}^{k-1} y_i$. 这些要求无疑对任意的 F_0 都成立. 这显然与不定积分相似.

11.18 求一个以算子 D 来表示 Δ^{-1} 的公式.

解 ☞ 结果 $e^D = 1 + \Delta$ 提示了

$$\Delta^{-1} = (e^D - 1)^{-1} = D^{-1}[D(e^D - 1)^{-1}],$$

其中 D^{-1} 为无穷积分算子, 它是 D 的逆. 从 Bernoulli 数的定义得

$$\begin{aligned}\Delta^{-1} &= D^{-1}\sum_{i=0}^{\infty}\frac{1}{i!}B_iD^i\\ &= D^{-1}\left(1 - \frac{1}{2}D + \frac{1}{12}D^2 - \frac{1}{720}D^4 + \cdots\right)\\ &= D^{-1} - \frac{1}{2} + \frac{1}{12}D - \frac{1}{720}D^3 + \cdots.\end{aligned}$$

像通常用不定积分那样(此处我们也有一个不定和)可以假定附加常数的存在.

11.19 通过算子导出 Euler-Maclaurin 公式.

解 ☞ 综合前二题的结果, 我们有

$$\begin{aligned}F_k &= \Delta^{-1}y_k = F_0 + \sum_{i=0}^{k-1} y_i\\ &= \left(D^{-1} - \frac{1}{2} + \frac{1}{12}D - \frac{1}{720}D^3 + \cdots\right)y_k.\end{aligned}$$

由第一个等式得

$$F_n - F_0 = \sum_{i=0}^{n-1} y_i,$$

而从第二个得

$$\begin{aligned}F_n - F_0 = \frac{1}{h}\int_{x_0}^{x_n} y(x)\mathrm{d}x - \frac{1}{2}(y_n - y_0) + \frac{h}{12}(y'_n - y'_0)\\ - \frac{h^3}{720}(y_n^{(3)} - y_0^{(3)}) + \cdots,\end{aligned}$$

所以最终得

$$\sum_{i=0}^{n-1} y_i = \frac{1}{h}\int_{x_0}^{x_n} y(x)\mathrm{d}x - \frac{1}{2}(y_n - y_0) + \frac{h}{12}(y'_n - y'_0) + \cdots,$$

这就是 Euler-Maclaurin 公式. 在这个推导中显然算子算术是支撑逻辑所必须的, 虽然它的来源可能有问题, 而且这样所得的级数通常是不收敛的, 但是其结果是有用的.

补 充 题

11.20 找出 $\sin x$ 及 $\cos x$ 在 $x_0 = 0$ 处的 n 次 Taylor 多项式.

11.21 对 $\sin x$ 与 $\cos x$ 以 Lagrange 形式表示误差项. 证明当 $n\to\infty$ 时对任意自变量 x, 这误差的极限为 0.

11.22 取什么样的 n 值 Taylor 多项式在 $0 < x < \pi/2$ 中逼近 $\sin x$ 准确到三位小数?

11.23 取什么样的 n 值才可以使 Taylor 多项式逼近 $\cos x$ 在 $0 < x < \pi/2$ 中准确到三位小数? 准确到 6 位呢?

11.24 把算子 Δ 表示成 D 的级数算子.

11.25 $\sinh x$ 及 $\cosh x$ 函数定义为

$$\sinh x = \frac{e^x - e^{-x}}{2}, \quad \cosh x = \frac{e^x + e^{-x}}{2}.$$

证明它们的 Taylor 级数为

$$\sinh x = \sum_{i=0}^{\infty} \frac{1}{(2i+1)!} x^{2i+1},$$

$$\cosh x = \sum_{i=0}^{\infty} \frac{1}{(2i)!} x^{2i}.$$

11.26 以算子算术表示

$$\delta = 2\sinh \frac{1}{2} D, \quad \mu = \cosh \frac{1}{2} D.$$

11.27 用二项式级数将 $\Delta = \frac{1}{2}\delta^2 + \delta\sqrt{1+\frac{1}{4}\delta^2}$ 表示为 δ 的幂级数直至 δ^7 项.

11.28 把题 11.13 和 11.27 的结果合在一起将 D 表示成 δ 的幂级数，可验证这些项直到 δ^7 为：

$$D = \delta - \frac{1^2}{2^2 3!}\delta^3 + \frac{1^2 \cdot 3^2}{2^4 5!}\delta^5 - \frac{1^2 \cdot 3^2 \cdot 5^2}{2^6 7!}\delta^7 + \cdots.$$

11.29 把题 11.28 的结果加以平方并按不同幂次的 δ 括在一起，可验证 D^2 的 Taylor 级数的这些项为

$$D^2 = \delta^2 - \frac{1}{12}\delta^4 + \frac{1}{90}\delta^6 - \frac{1}{560}\delta^8 + \frac{1}{3150}\delta^{10} - \cdots.$$

第十二章 插 值

历史地位

前面的章节几乎全都是基础性理论.现在要把理论用于若干方面.首先用于插值的古典问题.插值是用于估计函数值 $y(x)$的常用过程.自变量 x 在点 $x_0, \cdots, x_n$ 之间,而这些点上的值 $y_0, \cdots, y_n$ 是已知的.反插值就是简单地按反方向进行.表的加密是在点对 x_i 与 x_{i+1}之间作多值的系统插值以缩短数值表的间距,也许从 h 减至 $h/10$.预测是要求在自变量数据所落的区间外之 x 处对 $y(x)$的值作出估计.

所有这些运算在高速计算机问世之前尤其显得迫切,如今计算机能通过级数或别的非列表的途径来计算所有的常用函数值.本章的公式都冠有上世纪或更久远一些的杰出数学家的名字,那时函数表是必不可少的.它们的地位在我们的主题中是部分历史性的.令人感兴趣的是看到早期的计算障碍如何被越过,但重要的是注意到特殊函数表依然被造出来,使得这方面的某些工作继续发挥作用.

解法

插值方法着眼于将 $y(x)$代以某些易于计算的函数,通常是一个多项式,而所有的当中最简单的是直线.可将值 $y_0, \cdots, y_n$ 引入到我们的任何一个多项式中(Newton 公式,Everett 公式,…),随之它就成为一个插值算法,输出的就是 $y(x)$的逼近值.人们已经认识到同时用来自插值点两侧的数据是有意义的,能得到更好的结果或更简明的计算.Stirling、Bessel 和 Everett 等公式被鼓动其原因就在此,对其所含的误差的研究提供逻辑上的支撑.在表的两端点这做不到,应该用 Newton 向前和向后公式.不必要事先选定逼近多项式的次数,可简单地连续拟合表中的差直到适当的位数被认可为止.人们还认识到存在回报折返点,在那里结果变差了而不是被改进,并且这一点依赖于表值的精度.

另外一种是 Lagrange 方法,以多项式拟合数据不用有限差分.必须预先选定次数,但这方法具有补偿损失的优点.另一种不同的方法是 Aitken 方法,不要求自变量的表值为等距的或者在开头就要定多项式的次数.

密切多项式以及 Taylor 多项式在特殊情况下对插值问题也可以找到应用.

输入误差和算法误差

在所有这些应用中都存在输入误差和算法误差.它们对整个输出算数的影响只能估计到一定程度.习惯上确定为三种误差来源.

1. 输入误差产生于当给定的数据 $y_0, \cdots, y_n$ 不精确时,通常它们是实验值或计算所得.
2. 截断误差是指差 $y(x)-p(x)$,一旦我们决定采用某种多项式逼近,也就是对它的认可.先前所得到的误差为

$$y(x)-p(x)=\frac{\pi(x)}{(n+1)!}y^{(n+1)}(\xi).$$

   虽然 ξ 为未知量,但该公式眼前仍可以用来得到误差界.截断误差是一类算法误差.在预测问题中该误差可能是本质性的,因为在 $x_0, \cdots, x_n$ 所落的区间之外因子 $\pi(x)$变得特别地大.
3. 舍入误差则由于计算机运算是以固定的位数进行计算的,任何由乘、除所产生的多余位数将丢失.这是另一类算法误差.

题 解

12.1 预测缺少的二个 y_k 值.

$k=x_k$	0	1	2	3	4	5	6	7
y_k	1	2	4	8	15	26		

解 ☞ 这是一个简单的例子,但是它将起到提醒我们的作用,作为该应用之基础的是多项式逼近.计算某些差分

$$\begin{matrix} 1 & 2 & 4 & 7 & 11 \\ & 1 & 2 & 3 & 4 \\ & & 1 & 1 & 1 \end{matrix}$$

完全可以设想缺少的 y_k 可以为任何数,但是这些差分迹象强烈地指向一个二次多项式,暗示给定的6个值以及要预测的两个值都属于这样的一个多项式.接受这一点作为预测的基础,以至于无需再去找这个配置多项式.给三阶差分行加上两个1后,我们能很快地给二阶差分行补充一个5和一个6,16和22作为新的一阶差分,然后就预测得 $y_6=42$, $y_7=64$.这是在题6.12中所用的相同值,那儿所得到的配置多项式为三次的.

12.2 $y(x)=\sqrt{x}$ 的值列在表12.1中,舍入到4位小数,对于点 $x=1.00(0.01)1.06$(这表示点从1.00到1.06间隔为等距的取 $h=0.01$),计算差到 Δ^6 并解释它的意义.

解 ☞ 差也都列在表12.1中.

为简单起见,在差分记录中开头的那些零通常全被省略.在这个表中所有的差分均取到小数点后第四位.虽然平方根函数肯定不是线性的,但其一阶差分几乎都是常数,这暗示在整个列表的区间上取精度到4位小数时该函数可以用一个线性多项式精确地逼近.表值 Δ^2 最好看作是一个单位的舍入误差,而它对更高阶差分的影响遵从常用的二项式系数模式,如在题3.10中所观察到的那样.在这种情况下人们通常只要计算一阶差分.许多常用函数诸如 $\sqrt{x}$, $\log x$, $\sin x$ 等等都曾以这种方式列表,其自变量的间隔密度使得一阶差分几乎为常数,而且函数可以精确地用一个线性多项式来逼近.

表 12.1

x	$y(x)=\sqrt{x}$	Δ	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6
1.00	1.0000						
		50					
1.01	1.0050		0				
		50		−1			
1.02	1.0100		−1		2		
		49		1		−3	
1.03	1.0149		0		−1		4
		49		0		1	
1.04	1.0198		0		0		
		49		0			
1.05	1.0247		0				
		49					
1.06	1.0296						

12.3 应用 Newton 向前公式取 $n=1$ 对 $\sqrt{1.005}$ 进行插值.

解 ☞ Newton 公式可写成

$$p_k = y_0 + \binom{k}{1}\Delta y_0 + \binom{k}{2}\Delta^2 y_0 + \cdots + \binom{k}{n}\Delta^n y_0.$$

对线性逼近选 $n=1$，取 $k=\frac{x-x_0}{h}=\frac{1.005-1.00}{0.01}=\frac{1}{2}$ 我们得到

$$p_k = 1.0000 + \frac{1}{2}(0.0050) = 1.0025$$

这不能说是意外，因为我们用了一个线性配置多项式来匹配我们的 $y=\sqrt{x}$ 在 1.00 及 1.01 处的值，所以我们肯定可以预计到这个中间结果.

12.4 用更高次的多项式对题 12.3 进行插值会有怎样的效果?

解 ☞ 通过一个简单计算表明 Newton 公式的从二阶差分开始的几项近似地为 0.00001. 它们完全不会影响我们的结果.

12.5 $y(x)=\sqrt{x}$ 关于点 $x=1.00(0.05)1.30$ 舍入到 5 位小数的值列在表 12.2 中. 计算差分到 Δ^6 并说明它们的意义.

解 ☞ 差分列在表 12.2 中

表 12.2

x	$y(x)=\sqrt{x}$	Δ	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6
1.00	1.0000						
		2470					
1.05	1.02470		-59				
		2411		5			
1.10	1.04881		54		-1		
		2357		4		1	
1.15	1.07238		-50		-2		4
		2307		2		3	
1.20	1.09544		48		1		
		2259		3			
1.25	1.11803		-45				
		2214					
1.30	1.14017						

这儿误差的模式更为混乱一些，然而在最后 3 列中的 + 号及 − 号让人联想起题 3.10 及 3.11 中所产生的效应. 最好把这 3 列看成是误差效应，而不是把它看成计算平方根函数的有用信息.

12.6 使用题 12.5 中的数据对 $\sqrt{1.01}$ 进行插值.

解 ☞ 在表的顶部附近作插值用 Newton 向前公式是方便的. 在顶部表值 $x_0=1.00$ 处取 $k=0$，这种选择通常会导致项数的减少并使用多少项的决策几乎是自动的. 代入如在题 12.3 中所示的公式，取 $k=(x-x_0)/h=(1.01-1.00)/0.05=\frac{1}{5}$，我们得到

$$p_k = 1.00000 + \frac{1}{5}(0.02470) - \frac{2}{25}(-0.0059) + \frac{6}{125}(0.00005).$$

停止在这一项，因为它将不会影响第五位小数. 注意到这最后一项用的是题 12.5 中最高阶的差分，我们认为该差分对平方根的计算是有意义的. 我们没有无谓使用，已被预定只是误差效应的那些列，p_k 值简化为

$$p_k = 1.000000 + 0.004940 + 0.000048 + 0.000002 = 1.00499,$$

它准确到 5 位小数.（假如可能的话在计算过程中多加一位进行计算不失为一个好的想法，以此来控制在第一章中所描述过的"算法误差". 当然，在机器计算中，数字位数在任何情况下都是固定的，因此

这个注解不起作用.)

12.7　用题 12.5 中的数据对 $\sqrt{1.28}$ 进行插值.

解　☞ 这里用 Newton 向后公式是方便的,并且在题 12.6 中所做的大多数说明还有效. 在表底值 $x_0=1.30$ 处取 $k=0$,我们有 $k=(x-x_0)/h=(1.28-1.30)/0.05=-\frac{2}{3}$. 代入向后公式(题 7.9):

$$p_k=y_0+k\nabla y_0+\frac{k(k+1)}{2}\nabla^2 y_0+\frac{k(k+1)(k+2)}{3!}\nabla^3 y$$
$$+\cdots+\frac{k(k+1)\cdots(k+n-1)}{n!}\nabla^n y_0,$$

我们得到 $p_k=1.14017+\left(-\frac{2}{5}\right)(0.02214)+\left(-\frac{3}{25}\right)(-0.00045)$

$$+\left(-\frac{8}{125}\right)(0.00003)$$
$$=1.140170-0.008856+0.000054-0.000002$$
$$=1.13137.$$

它准确到 5 位小数.

12.8　前面二题处理的是插值问题的特殊情况,在靠近表的顶部或靠近底部进行工作. 本题是在插值点两侧的数据都可使用的更为典型的问题. 用题 12.5 中的数据对 $\sqrt{1.12}$ 进行插值.

解　☞ 现在使用中心差分公式是方便的,因为它使得能容易地采用在两侧大约相同数目的数据. 在题 12.15 中我们将会看到它还趋向于保持小的截断误差. Everett 公式将被使用.

$$p_k=\binom{k}{1}y_1+\binom{k+1}{3}\delta^2 y_1+\binom{k+2}{5}\delta^4 y_1+\cdots-\binom{k-1}{1}y_0$$
$$-\binom{k}{3}\delta^2 y_0-\binom{k+1}{5}\delta^4 y_0-\cdots,$$

这里略去了更高阶的项,因为在这个问题中已用不到它们. 在 $x_0=1.10$ 处选 $k=0$,我们有 $k=(x-x_0)/h=(1.12-1.10)/0.05=2/5$. 将它代入 Everett 公式

$$p_k=\left(\frac{2}{5}\right)(1.07238)+\left(-\frac{7}{125}\right)(-0.00050)+\left(\frac{168}{5^6}\right)(-0.00002)$$
$$-\left(-\frac{3}{5}\right)(1.04881)-\left(\frac{8}{125}\right)(-0.00054)-\left(-\frac{182}{5^6}\right)(-0.00001)$$
$$=0.428952+0.000028+0.629286+0.000035.$$

最高两项不起作用(正如所料,由于它们是从误差效应引出来的项). 最后 $p_k=1.05830$,它准确到 5 位小数. 注意,这三个由表 12.2 做成的插值都基于 3 次配置多项式.

12.9　实验室的新雇员被要求在国家标准局应用数学业书的表 NBS - AMS52 中寻求 $y(0.3333)$ 之值. 在这大部头书卷的某页上他获得丰富的信息,其中很小部分就列在表 12.3 中. 对所需的插值应用 Everett 公式.

表 12.3

x	$y(x)$	δ^2
0.31	0.1223 4609	2392
0.32	0.1266 9105	2378
0.33	0.1310 5979	2365
0.34	0.1354 5218	2349
0.35	0.1398 6806	2335

解 在 $x_0=0.33$ 处取 $k=0$* 我们有 $k=(x-x_0)/h=(0.3333-0.33)/0.01=0.33$. 将直到二阶差分的 Everett 公式写成如下形式:

$$p_k = ky_1+(1-k)y_0+E_1\delta^2y_1-E_0\delta^2y_0,$$

其中 $E_1=\binom{k+1}{3}$ 和 $E_0=\binom{k}{3}$. 插值公式可以在表中找到所有有用的成分. 对 $k=0.33$, 我们得到 $E_1=-0.0490105, E_0=0.0615395$. 于是

$$\begin{aligned}p_k &= (0.33)(0.13545218)+(0.67)(0.13105979)\\ &\quad +(-0.0490105)(0.00002349)-(0.0615395)(0.00002365)\\ &=0.13250667\end{aligned}$$

这表就是为 Everett 公式而准备的.

12.10 应用 Lagrange 公式由表 12.2 中的数据来得到 $\sqrt{1.12}$.

解 Lagrange 公式并不要求等距节点. 当然它可以应用到这种点看作特殊情况, 但是有困难. 配置多项式的次数必须从一开始便选定. 用 Newton, Everett 或其他差分公式时, 其次数可通过计算它们的项来决定, 后面的每一项是对已经累加起来的项的一个附加校正, 因此计算进行到所算的项不再起校正作用为止. 然而对 Lagrange 公式而言改变次数则意味着所有的项都要重新计算. 在表 12.2 中十分明显用一个三次多项式是恰当的. 据此, 从选 $x_0=1.05,\cdots,x_3=1.20$ 开始我们的工作, 并把它们代入

$$\begin{aligned}p &= \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}y_0+\frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}y_1\\ &\quad +\frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)}y_2+\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}y_3,\end{aligned}$$

得

$$\begin{aligned}p &= \frac{-8}{125}(1.02470)+\frac{84}{125}(1.04881)+\frac{56}{125}(1.07238)\\ &\quad +\frac{-7}{125}(1.09544)=1.05830.\end{aligned}$$

该结果与题 12.8 中的结果相一致.

12.11 反插值问题是将 x_k 与 y_k 的角色对换. 我们可以把 y_k 看成自变量, x_k 看成是函数值. 显见新的自变量通常其间隔不是等距的. 给定 $\sqrt{x}=1.05$ 用表 12.2 中的数值来求 x.

解 由于我们可以容易地通过一个简单的乘法得到 $x=(1.05)^2=1.1025$, 这可直接了当地充当我们所用算法的另一种"测试"("test case"). 由于它用的是非等距节点, 所以我们用 Lagrange 公式, 将 x 与 y 的角色对换,

$$\begin{aligned}p &= \frac{(y-y_1)(y-y_2)(y-y_3)}{(y_0-y_1)(y_0-y_2)(y_0-y_3)}x_0+\frac{(y-y_0)(y-y_2)(y-y_3)}{(y_1-y_0)(y_1-y_2)(y_1-y_3)}x_1\\ &\quad +\frac{(y-y_0)(y-y_1)(y-y_3)}{(y_2-y_0)(y_2-y_1)(y_2-y_3)}x_2+\frac{(y-y_0)(y-y_1)(y-y_2)}{(y_3-y_0)(y_3-y_1)(y_3-y_2)}x_3\end{aligned}$$

以用在题 12.10 中相同的四个点作为 x_k, y_k, 它就变成

$$\begin{aligned}p &= (-0.014882)1.05+(0.97095)1.10+(0.052790)1.15\\ &\quad +(-0.008858)1.20=1.1025,\end{aligned}$$

正如所期望的那样.

12.12 将 Everett 公式应用于适才解决了的反插值问题上.

解 由于 Eeverett 公式要求等距节点, 我们将 x 与 y 回到他们原来的角色. 将 Everett 公式写成

$$1.05=k(1.07238)+\binom{k+1}{3}(-0.00050)+\binom{k+2}{5}(-0.00002)$$

* 译注: 此处原文误为 $x=0$.

$$+(1-k)(1.04881)-\binom{k}{3}(-0.00054)-\binom{k+1}{5}(-0.00001),$$

我们有了一个关于 k 的5次多项式方程.这是一个在以后章节中将广泛探讨的问题.这里可以用一个简单的迭代过程.首先将所有差分丢在一边,通过解

$$1.05=k(1.07238)+(1-k)(1.04881)$$

来得到一个首次逼次.这个线性反插值的结果是 $k=0.0505$.将该值代入 δ^2 项,继续略去 δ^4 次项,从

$$1.05=k(1.07238)+\binom{1.0505}{3}(-0.00050)+(1-k)(1.04881)$$

$$-\binom{0.0505}{3}(0.00054)$$

得到一个新的逼近值,它是 $k=0.0501$.将这个值同时代入 δ^2 及 δ^4 项得 $k=0.0500$.再将它代入 δ^2 及 δ^4 项,最后得到的 k 值还是它自己,我们就到此为止.相应的取四位小数的 x 值为 1.1025

12.13 在表12.2中对 $\sqrt{1.125}$ 及 $\sqrt{1.175}$ 进行插值.

解 ☞ 对于这些处在列表节点中间的节点,Bessel 公式具有强烈的吸引力,首先在 $x_0=1.10$ 处选择 $k=0$,致使 $k=(1.125-1.10)/0.05=\frac{1}{2}$.Bessel 公式(题7.25)为

$$p_k=\mu y_{1/2}+\binom{k}{2}\mu\delta^2 y_{1/2}+\binom{k+1}{4}\mu\delta^4 y_{1/2},$$

这里我们取到4次项为止.奇次项由于因子 $(k-\frac{1}{2})$ 而全部消失.代入得

$$p_k=1.06060+\left(-\frac{1}{8}\right)(-0.00052)+\left(\frac{3}{128}\right)(-0.000015)$$

$$=1.06066,$$

δ^4 项还是没有任何贡献.类似地在第二种情况中,现在在 $x_0=1.15$ 处取 $k=0$,我们再一次有 $k=1/2$ 并得到 $p_k=1.08397$.通过寻找所有这类中点值,表的大小可以增大一倍,这是表加密的一种特殊情况.

12.14 在使用配置多项式 $p(x)$ 来计算对一个函数 $y(x)$ 的逼近时,我们将 $y(x)-p(x)$ 称为截断误差.对表12.1中的插值估计该误差.

解 ☞ 当逼近多项式为 n 次时,在第2章中导出的配置多项式的截断误差公式为

$$y(x)-p(x)=\frac{\pi(x)}{(n+1)!}y^{(n+1)}(\xi),$$

对于表12.1我们发现 $n=1$ 为恰当的.配置点可以记为 x_0 及 x_1,导出关于线性插值的误差估计:

$$y(x)-p(x)=\frac{(x-x_0)(x-x_1)}{2}y^{(2)}(\xi)$$

$$=\frac{k(k-1)}{2}h^2y^{(2)}(\xi).$$

由于 $h=0.1$ 以及 $y^{(2)}(x)=-\frac{1}{4}x^{-3/2}$,我们有

$$|y(x)-p(x)|\leqslant\frac{k(k-1)}{8}(0.0001)^*$$

对任何内插通过选择 x_0 我们可安排 k 在0与1之间,对这种安排,二次式 $k(k-1)$ 在中点 $k=\frac{1}{2}$ 处有最大幅度 $\frac{1}{4}$(参看图12.1).这就使我们完成了截断误差的估计:

$$|y(x)-p(x)|\leqslant\frac{1}{32}(0.0001).$$

---

* 译注:严格地说,这个不等式是有缺陷的,因为当 $0<k<1$ 时,$k(k-1)$ 是负的;这个误差估计式的正确写法应为

$$|y(x)-p(x)|\leqslant\frac{|k(k-1)|}{8}(0.0001).$$

并由此我们发现它不会影响到第4位小数.表12.1就是为线性插值而准备的.区间选成 $h=0.01$ 是为了保持截断误差能达到如此之小.

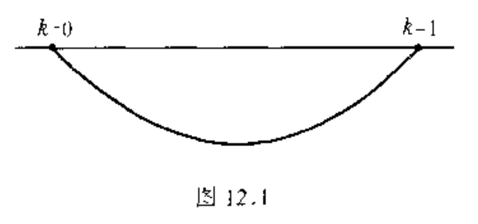


图 12.1

12.15 估计我们的表 12.2 进行计算时的截断误差.

解 ☞ 这里我们大部分使用的是三次多项式的 Everett 公式.对于其他的三次公式得出同样的误差估计.假设配置点 x_{-1},x_0,x_1 及 x_2 为等距的,

$$y(x)-p(x)=\frac{(x-x_{-1})(x-x_0)(x-x_1)(x-x_2)}{4!}y^{(4)}(\xi)$$
$$=\frac{(k+1)k(k-1)(k-2)h^4y^{(4)}(\xi)}{24}.$$

多项式 $(k+1)k(k-1)(k-2)$ 具有图 12.2 的一般形式,在区间 $-1<k<2$ 之外它迅速地升向上方.在 $0<k<1$ 之内它不超过 $\frac{9}{16}$,并且这是插值的适合部分.现在关于三次插值的最大误差有

$$|y(x)-p(x)|\leqslant\frac{9}{16}\cdot\frac{1}{24}h^4|y^{(4)}(\xi)|=\frac{3}{128}h^4|y^{(4)}(\xi)|.$$

对这个例子来说,$h=0.05$ 而 $y^{(4)}(x)=-\frac{15}{16}x^{-7/2}$,因此 $|y(x)-p(x)|\leqslant\frac{1}{64}(0.00005)$,所以截断误差不影响到我们计算的第五位小数.

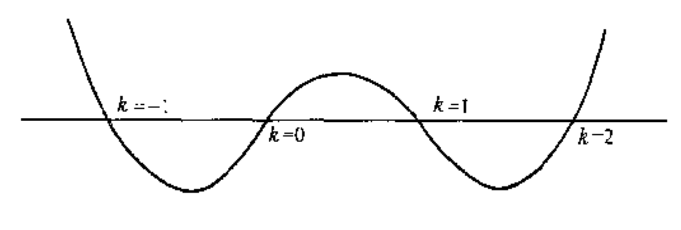


图 12.2

12.16 在 $\sqrt{x}$ 的表中用一个三次公式,区间长度 h 取多大时还能给出5位的精确度?(假设 $1\leqslant x$)

解 ☞ 这类问题对造表者而言自然是感兴趣的.我们的截断误差公式可以写成

$$|y(x)-p(x)|\leqslant\frac{9}{16}h^4\left(\frac{15}{16}\right)\left(\frac{1}{24}\right),$$

要它保持小于 0.000005 就要求 $h^4<0.000228$ 或者说很接近于 $h<\frac{1}{8}$.这比表 12.1 中所用的 $h=0.05$ 略大一些.但是进入我们计算中的其它误差就要付出不稳的代价了.

12.17 前一题表明,假如用 Everett 三次多项式来插值的话,表 12.2 的长度可以缩减一半.找出这个 Everett 公式中所需的二阶差分.

解 ☞ 这结果列在表 12.4 中,表中一阶差分可以忽略.

表 12.4

x_k	y_k	δ	δ^2
1.00	1.00000		
		4881	
1.10	1.04881		−217
		4664	
1.20	1.09544		−191
		4473	
1.30	1.14017		

12.18 用表 12.4 来插值 $y(1.15)$.

解 ☞ 取 Everett 公式及 $k=\frac{1}{2}$,

$$p_k=\frac{1}{2}(1.09544)-\frac{1}{16}(-0.00191)+\frac{1}{2}(1.04881)$$
$$-\frac{1}{16}(-0.00217)=1.07238,$$

如在表 12.2 中所列.这个例子进一步证实题 12.16.

12.19 对一个 5 次公式估计截断误差.

解 ☞ 假设配置点为等距的并且与 Everett 公中一样,在 $k=-2,-1,\cdots,3$ 处(位置其实是不重要的).

$$y(x)-p(x)=\frac{\pi(x)}{(n+1)!}y^{(n+1)}(\xi)$$
$$=\frac{(k+2)(k+1)k(k-1)(k-2)(k-3)}{720}h^6y^{(6)}(\xi),$$

分子因子当 $0<k<1$ 时在 $k=\frac{1}{2}$ 处取极大绝对值 $\frac{225}{64}$,易证其结果为

$$|y(x)-p(x)|\leqslant\frac{1}{720}\cdot\frac{225}{64}\cdot h^6|y^{(6)}(\xi)|.$$

12.20 对于函数 $y(x)=\sqrt{x}$ 且 $1\leqslant x$,假如在插值时使用 Everett5 次公式,与 5 位精度相容的 h 应取多大?

解 ☞ 对于这个函数而言,$y^{(6)}(x)=-\frac{945}{64}x^{-11/2}\leqslant\frac{945}{64}$,将这个结果代入上题中并要求 5 位精度:

$$\frac{1}{720}\cdot\frac{225}{64}\cdot h^6\cdot\frac{945}{64}\leqslant 0.000005,$$

近似地导出 $h\leqslant\frac{1}{5}$.自然地,5 次插值所允许的区间比 3 次插值的大.

12.21 对于函数 $y(x)=\sin x$,假如用 5 次 Everett 公式进行插值,h 应取多大才与 5 位精度相容?

解 ☞ 对该函数而言,$y^{(6)}(x)$ 的绝对值界为 1,所以我们需要 $\frac{1}{720}\cdot\frac{225}{64}\cdot h^6\leqslant 0.000005$,由此 $h\leqslant 0.317$,它等价于区间大小为 18°,这意味着,除 sin0° 及 sin90° 外只需 4 个正弦函数值就可以覆盖整个基本区间!

12.22 在用我们的关于配置多项式的公式时,其第二种误差来源(第一种来自截断误差)为在数据值中存在不精确性.例如数 y_k,假如它是通过物理测量所得到的,由于仪器本身的局限性就含有不精确性.倘若由计算所得,它就可能包含有舍入误差.证明在线性插值过程中这类误差不会被放大.

证 ☞ 线性多项式可以写成 Lagrange 形式

$$p=ky_1+(1-k)y_0,$$

其中 y_k 如通常那样为实际数据值.假设这些值是不精确的,用 Y_1 及 Y_0 表精确的然而是未知的值,我们可以记

$$Y_0=y_0+e_0,\quad Y_1=y_1+e_1,$$

其中 e_0 和 e_1 为误差.因此所要求的精确值为

$$P=kY_1+(1-k)Y_0.$$

造成与我们计算结果之间的误差为

$$P-p=ke_1+(1-k)e_0.$$

假如误差 e_k 的大小不超过 E,则

$$|P-p|\leqslant kE+(1-k)E=E.$$

当 $0<k<1$ 时计算值 p 的误差不会超过最大的数据误差,没有出现误差的放大.

12.23 估计由三次插值带来的数据不精确性的放大率.

解 ☞ 还是使用 Lagrange 形式,但是假设在 $k=-1,0,1,2$ 处自变量值为等距的,三次多项式可

以写作

$$p = -\frac{k(k-1)(k-2)}{6}y_{-1} + \frac{(k+1)(k-1)(k-2)}{2}y_0$$
$$+\frac{(k+1)k(k-2)}{-2}y_1 + \frac{(k+1)k(k-1)}{6}y_2.$$

正如在题 12.22 中那样，我们令 $Y_k = y_k + e_k$ 以 Y_k 表示精确数据值，假如仍以 P 表所要求的精确值，则误差为

$$P - p = -\frac{k(k-1)(k-2)}{6}e_{-1} + \frac{(k+1)(k-1)(k-2)}{2}e_0$$
$$+\frac{(k+1)k(k-2)}{-2}e_1 + \frac{(k+1)k(k-1)}{6}e_2.$$

注意当 $0<k<1$ 时误差 e_{-1} 及 e_2 有负系数，而另二个有正系数. 这意味着如果误差的大小不超过 E，则

$$|P-p| \leqslant E\left[\frac{k(k-1)(k-2)}{6} + \frac{(k+1)(k-1)(k-2)}{2}\right.$$
$$\left.+\frac{(k+1)k(k-2)}{-2} + \frac{(k+1)k(k-1)}{-6}\right].$$

它简化为

$$|P-p| \leqslant (-k^2+k+1)E = m_kE.$$

毫无疑问，二次放大因子 m_k 在 $k=\frac{1}{2}$ 处将取它的极大值(见图 12.3)，于是 $|P-p| \leqslant \frac{5}{4}E$. 数据误差 E 至多被放大 $\frac{5}{4}$ 倍. 当然这是一种悲观的估计. 在某种情况下误差甚至可以相互抵消，致使计算所得出的 p 值较数据 y_k 更为精确.

12.24 在插值过程中还有什么其它的误差来源呢？

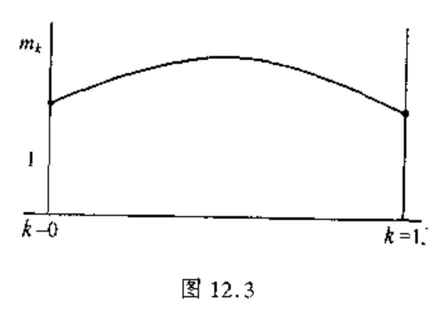


图 12.3

解 ☞ 一种非常重要的误差来源必需予以关注，那就是在算法的整个计算过程中一直都会产生的舍入误差，它常常完全不受人们的控制. 只能用有限位数进行计算，这是无法避免的. 我们所用的各种公式(即使它们精确地表示同一配置多项式)将以不同的方法来处理所涉及的数据. 换言之，它们代表不同的算法，这类公式接受同样的输入误差(数据不精确性)而且可以有相同的截断误差，然而算法舍入误差的发展过程依然不同.

12.25 描述 Taylor 级数是怎样地用于插值的.

解 ☞ 考虑函数 $y=e^x$. 然而 Taylor 级数为

$$e^{x+t} = e^x \cdot e^t = e^x\left(1 + t + \frac{1}{2}t^2 + \cdots\right).$$

假设因子 e^x 为已知的. 将该级数在 t^2 项后截断，这表示误差(在圆括号内)最多为 $\frac{1}{6}(h/2)^3$，其中 h 为表中自变量值的间隔，这就设想插值通常基于最靠近的表值. 假如 $h=0.05$，这个误差就是 $\left(\frac{125}{48}\right)10^{-6}$ 或是 $(2.6)10^{-6}$. 它说明，若在 t^2 项处停止，在计算 e^{x+t} 的值时可以精确到 5 位十进制数(不是指小数位数). 例如，用表 12.5 中的数据，$e^{2.718}$ 的插值进行如下，取 $t=0.018$，$1+t+\frac{1}{2}t^2=1.01816$ 及

$$e^{2.718} = e^{2.70}(1.01816) = (14.880)(1.01816) = 15.150.$$

它的全部 5 位都是正确的，我们的配置多项式也会产生同样的结果.

表 12.5

x	2.60	2.65	2.70	2.75	2.80
$y=e^x$	13.464	14.154	14.880	15.643	16.445

12.26 Taylor 级数插值怎样用在函数 $y(x)=\sin x$ 上的?

解 ☞ 由于 $\sin x$ 及 $\cos x$ 通常都是一起列表的,我们可以将它表示成

$$\sin(x\pm t)=\sin x\pm t\cos x-\frac{1}{2}t^2\sin x.$$

自然,这里的 t 是以弧度来度量的.假如表的间隔是 $h=0.0001$,在表 NBS-AMS36 中正是如此,表 12.6 是从中摘录出来的一小段,则上面的公式将给出 9 位精度,因为$\frac{1}{6}(h/2)^3$ 已落在 12 位之外.

表 12.6

x	$\sin x$	$\cos x$
1.0000	0.8414 70985	0.5403 02306
1.0001	0.8415 25011	0.5402 18156
1.0002	0.8415 79028	0.5401 34001
1.0003	0.8416 33038	0.5400 49840

12.27 以 Taylor 级数插值计算 sin1.00005.

解 ☞ 以 $x=1$ 及 $t=0.00005$

$$\sin 1.00005=0.8414\ 70985+(0.00005)(0.5403\ 02306)$$
$$-\left(\frac{1}{8}\right)(10^{-8})(0.8414\ 70985)=0.8414\ 97999.$$

12.28 应用 Newton 向后公式并用表 12.2 来预测 $\sqrt{1.32}$.

解 ☞ 在 $x_0=1.30$ 处,取 $k=0$ 我们得到 $k=(1.32-1.30)/0.05=0.4$.代入 Newton 公式得

$$p=1.14017+(0.4)(0.02214)+(0.28)(-0.00045)$$
$$+(0.224)(0.00003)=1.14891,$$

到目前为止它是准确的.Newton 向后公式看来是对这类预测问题的自然选择,因为能为这个公式提供最多的有用差分,人们可以引入差分项直到它们对保留的小数位数不再有供献为止.这就使在计算机进行的过程中来选定逼近多项式的次数成为可能.

12.29 分析在预测中的截断误差.

解 ☞ 配置多项式的截断误差可以表示为

$$\frac{k(k+1)\cdots(k+n)}{(n+1)!}h^{n+1}y^{(n+1)}(\xi),$$

其中配置点对应 $k=0,-1,\cdots,-n$,如同用 Newton 向后公式的那种情况.对预测来说 k 为正的 分子上的因子随 k 的增长迅速增大,对于大的 k 而言增得更快,如图 12.4 所示.这说明截断误差超过某个点后就失去控制,远离表端处作预测是危险的,正如所料.配置多项式的截断误差在配置点之间振动,一旦在这些点的区间之外就变得不可收拾了.

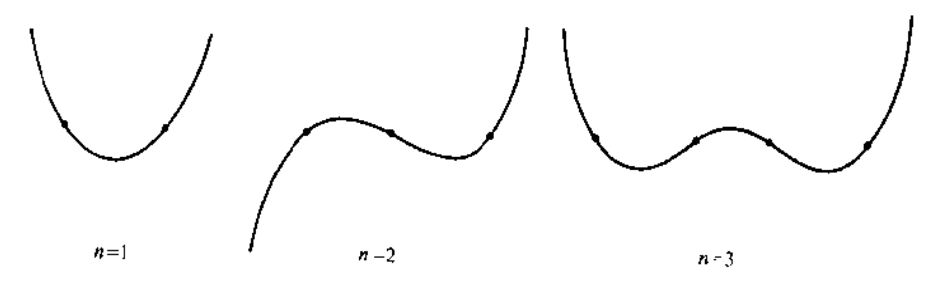


图 12.4

12.30 由表 12.2 的数据预测 $\sqrt{1.50}$.

解 ☞ 当 $k=(1.50-1.30)/0.05=4$,

$$p=1.14017-(4)(0.02214)+(10)(-0.00045)+(20)(0.00003)=1.22483,$$

而准确值为 1.22474,还应注意高阶差分项,我们相信在任何情况下它都是误差效应,由于它们是正的只会使结果变坏.

补　充　题

12.31 用线性插值从表 12.1 的数据中得到 $\sqrt{1.012}$ 和 $\sqrt{1.017}$,到 4 位小数.二阶差分项是否会影响该结果?更高阶的项又怎样呢?

12.32 用线性插值从表 12.1 的数据中得到 $\sqrt{1.059}$.注意到假如用 Newton 向前公式(在 $x=1.05$ 处取 $k=0$),在这种情况下二阶差分不起作用.

12.33 用表 12.2 对 $\sqrt{1.03}$ 进行插值.

12.34 用表 12.2 对 $\sqrt{1.26}$ 进行插值.

12.35 应用 Stirling 公式从表 12.2 的数据中求得 $\sqrt{1.12}$.这个结果与题 12.8 的结果是否一致?

12.36 对表 12.3 应用 Everett 公式求 $y(0.315)$.

12.37 应用 Lagrange 公式并用标准误差函数列,在下面的某些值对 $y(1.50)$ 进行插值.$y(x)=e^{-x^2/2}\sqrt{2\pi}$.

x_k	1.00	1.20	1.40	1.60	1.80	2.00
y_k	0.2420	0.1942	0.1497	0.1109	0.0790	0.0540

准确值为 0.1295.

12.38 用 Lagrange 公式以题 12.37 中的数据对相应于 $y=0.1300$ 的 x 进行反插值.

12.39 应用题 12.12 中的方法求题 12.38 的反插值.

12.40 对题 12.37 中的数据应用 Bessel 公式求 $y(1.30)$, $y(1.50)$ 及 $y(1.70)$.

12.41 在函数 $y(x)=\sin x$ 的取四位小数的表中,与线性插值相容的最大区间长度 h 为何?(保持截断误差低于 0.00005.)

12.42 在 5 位的 $y(x)=\sin x$ 的表中,与线性插值相容的最大区间长度 h 为何?对照常用的正弦函数表来检查这些估计值.

12.43 假如用三次多项式进行插值,而不是线性插值,在一个 $y(x)=\sin x$ 的 4 位小数表中,最大可以用的区间 h 为何?在一个 5 位的表中又如何呢?

12.44 用 Newton 公式作二次逼近时,函数 $k(k-1)(k-2)$ 出现在截断误差估计中.证明这个函数具有在图 12.5 中所示的形式,并且当 $0<k<2$ 时其绝对值不超过 $2\sqrt{3}/9$.

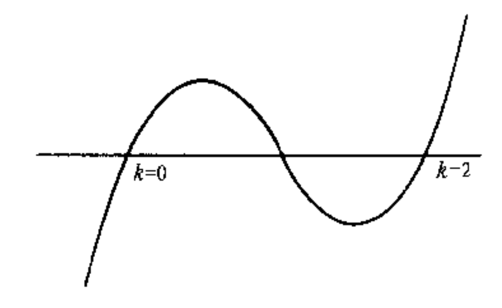


图 12.5

12.45 函数 $k(k^2-1)(k^2-4)$ 出现在 Stirling 公式的截断误差中.对 $-2<k<2$ 画出简图.估计它在区间 $-1/4<k<1/4$ 内的最大绝对值,通常限在这区间内使用该公式.

12.46 证明多项式

$$k(k^2-1)(k^2-4)\quad 和\quad k(k^2-1)(k^2-4)(k^2-9)$$

的相对极大与极小值之大小随它们离区间 $-1<k<1$ 之距离的加大而增大. 这些多项式出现在 Stirling 公式的截断误差中. 这隐含着在配置区域的中部这公式最为精确.

12.47 证明多项式

$$(k+1)k(k-1)(k-2) \text{ 和 } (k+2)(k+1)k(k-1)(k-2)(k-3)$$

的相对极大与极小值随着离区间 $0<k<1$ 之距离的加大而增大. 这些多项式出现在 Everett 或 Bessel 公式的截断误差中, 这隐含着这些公式在这整个中心区间内最精确.

12.48 以 Everett5 次公式进行插值, 假如函数是 $y(x)=\log x$ 而且要求 5 位精度, 与之相容的区间大小为何?

12.49 估计由二次插值造成的数据不精确性的放大, 按题 12.22 及 12.23 中的论述, 取 $0<k<1$

12.50 估计由四次插值造成的数据的不精确性的放大, 仍是对 $0<k<1$

12.51 从表 12.5 中的数据应用 Stirling 公式计算 $y(2.718)$.

12.52 由表 12.6 提供的数据计算 sin1.00015.

12.53 证明 Taylor 级数插值

$$\log(x+t)=\log x+\log\left(1+\frac{t}{x}\right)$$
$$=\log x+\frac{t}{x}-\frac{t^2}{2x^2}+\cdots$$

可以在 t^2 项后截断对 $1<x$ 具有 6 位小数精度, 假定表间隔 $h=0.01$.

12.54 从表 12.2 中的数据, 以 Newton 向后公式预测 $\sqrt{1.35}$, $\sqrt{1.40}$ 及 $\sqrt{1.45}$.

12.55 从表 12.4 中的数据预测 $\sqrt{1.40}$ 及 $\sqrt{1.50}$.

12.56 分析题 6.14 中二次多项式的误差, 证明误差在 $x=-3$ 以及在配置点上误差为零. 怎样能用我们的配置误差公式 $\pi(x)y^{(3)}(\xi)/3!$ 来解释它?

12.57 在题 6.15 中怎样能用误差公式 $\pi(x)y^{(4)}(\xi)/4!$ 来解释在 $x=4$ 处的零误差.

12.58 用题 10.15 的结果估计缺项 $y'(1)$.

12.59 用题 10.16 的结果估计缺项 $y''(1)$.

12.60 用题 10.17 的结果估计缺项 $y'(0)$ 和 $y'(1)$.

第十三章　数值微分

近似导数

函数 $y(x)$ 的近似导数可简单地用其近似多项式的各阶导数 $p', p^{(2)}, p^{(3)}$ 来代替 y', $y^{(2)}, y^{(3)}\cdots$. 通过配置多项式可导出大量这类有用的公式. 三个最著名的公式

$$y'(x) \approx \frac{y(x+h)-y(x)}{h}, \quad y'(x) \approx \frac{y(x+h)-y(x-h)}{2h},$$

$$y'(x) \approx \frac{y(x)-y(x-h)}{h}$$

分别由微商的 Newton 向前公式, Stirling 公式和 Newton 向后公式在每一种情况下仅用一项而得到的. 简单地使用较多项就可得到更为复杂的公式. 因此由 Newton 向前公式可得

$$y'(x) \approx \frac{1}{h}\left[\Delta y_0 + \left(k - \frac{1}{2}\right)\Delta^2 y_0 + \frac{3k^2-6k+2}{6}\Delta^3 y_0 + \cdots\right].$$

而通过微分商 Stirling 公式可得

$$y'(x) \approx \frac{1}{h}\left(\delta\mu y_0 + k\delta^2 y_0 + \frac{3k^2-1}{6}\delta^3\mu y_0 + \cdots\right).$$

由其他的配置公式产生类似的近似. 关于二阶导数, 一个著名的结果由 Stirling 公式得到的

$$y^{(2)}(x) \approx \frac{1}{h^2}\left(\delta^2 y_0 + k\delta^3\mu y_0 + \frac{6k^2-1}{12}\delta^4 y_0 + \cdots\right).$$

只保留第一项便有熟悉的公式

$$y^{(2)}(x) \approx \frac{y(x+h)-2y(x)+y(x-h)}{h^2}.$$

近似微分中的误差来源

对试验情形的研究提示, 由配置多项式得到的近似导数看来是有怀疑的, 除非有非常精确的数据可用. 即使这样精确度随导数阶数的增加而减少.

根本困难是当 $y'(x)-p'(x)$ 很大时, $y(x)-p(x)$ 可能很小, 用几何语言表示, 即两条曲线可以很接近, 但其斜率却相差很大. 也可能出现所有其他熟悉的误差来源, 包括 y_i 值的输入误差, 诸如 $y'-p'$, $y^{(2)}-p^{(2)}$ 的截断误差等等以及固有的舍入误差.

主要的误差来源是输入误差本身. 这些是关键的, 因为即使它们很小时, 因为算法会把它们扩大很多. 这种放大的决定因素是出现在公式中的 h 的倒数次幂. 乘上精确值和误差, 它们混合在一起给出数据 y_i. 有时可最优选择区间长 h, 因为截断误差与 h 成正比而输入误差的放大与 h 成反比. 一般的算法可使其两者的结合为最小.

基于配置多项式的近似导数, 应预先考虑大的误差. 只要有可能就应得到误差界. 求近似微分的其他方法可以是基于最小二乘方法及最小—最大方法得到的多项式(见二十一章和二十二章), 而不是配置多项式. 因为这些方法还对所给数据进行光滑. 所以它们通常更令人满意. 三角近似法(二十四章)提供了另一种方法.

题　　解

13.1　微分 Newton 向前公式

$$p_k = y_0 + \binom{k}{1}\Delta y_0 + \binom{k}{2}\Delta^2 y_0 + \binom{k}{3}\Delta^3 y_0 + \binom{k}{4}\Delta^4 y_0 + \cdots$$

解 ☞ 可以用 Stirling 数把阶乘表成幂,于是通过简单的计算就得出关于 k 的导数用算子 D 继续表示这样的导数,$Dp_k, D^2p_k, \cdots$,我们用熟悉的 $x = x_0 + kh$ 得到关于自变量 x 的导数

$$p'(x) = \frac{Dp_k}{h}, \qquad p^{(2)}(x) = \frac{D^2p_k}{h^2}, \cdots$$

结果是

$$p'(x) = \frac{1}{h}\left[\Delta y_0 + \left(k - \frac{1}{2}\right)\Delta^2 y_0 + \frac{3k^2 - 6k + 2}{6}\Delta^3 y_0 + \frac{2k^3 - 9k^2 + 11k - 3}{12}\Delta^4 y_0 + \cdots\right],$$

$$p^{(2)}(x) = \frac{1}{h^2}\left\{\Delta^2 y_0 + (k - 1)\Delta^3 y_0 + \frac{6k^2 - 18k + 11}{12}\Delta^4 y_0 + \cdots\right\},$$

$$p^{(3)}(x) = \frac{1}{h^3}\left\{\Delta^3 y_0 + \frac{2k - 3}{2}\Delta^4 y_0 + \cdots\right\},$$

$$p^{(4)}(x) = \frac{1}{h^4}(\Delta^4 y_0 + \cdots), \text{等等}.$$

13.2 用题 13.1 中的公式,根据表 13.1 的数据来求 $p'(1)$, $p^{(2)}(1)$ 和 $p^{(3)}(1)$.(这是与表 12.2 相同的直到三阶的差分.因误差的影响没有写出其余的差分.这里为方便起见复制了该表)

表 13.1

x	$y(x) = \sqrt{x}$			
1.00	1.00000			
		2470		
1.05	1.02470		−59	
		2411		5
1.10	1.04881		−54	
		2357		4
1.15	1.07238		−50	
		2307		2
1.20	1.09544		−48	
		2259		3
1.25	1.11803		−45	
		2214		
1.30	1.14017			

解 ☞ 取 $h = 0.05, k = 0$ 在 $x_0 = 1.00$,由我们的公式得到

$$p'(1) = 20(0.02470 + 0.000295 + 0.00017) = 0.50024$$

$$p^{(2)}(1) = 400(-0.00059 - 0.00005) = -0.256$$

$$p^{(3)}(1) = 8000(0.00005) = 0.4$$

因为 $y(x) = \sqrt{x}$,准确值是 $y'(1) = \frac{1}{2}, y^{(2)}(1) = -\frac{1}{4}, y^{(3)}(1) = \frac{3}{8}$.

尽管输入数据精确到五位小数,我们发现 $p'(1)$ 仅精确到三位,$p^{(2)}(1)$ 两位小数也不完全精确,$p^{(3)}(1)$ 仅精确到一位.显然,算法的误差是明显的.

13.3 微分 Stirling 公式

$$p_k = y_0 + \binom{k}{1}\delta\mu y_0 + \frac{k}{2}\binom{k}{1}\delta^2 y_0 + \binom{k+1}{3}\delta^3\mu y_0 + \frac{k}{4}\binom{k+1}{3}\delta^4 y_0 + \cdots$$

解 ☞ 如同在题 13.1 中那样,我们得到

$$p'(x) = \frac{1}{h}\left(\delta\mu y_0 + k\delta^2 y_0 + \frac{3k^2 - 1}{6}\delta^3\mu y_0 + \frac{2k^3 - k}{12}\delta^4 y_0 + \cdots\right),$$

$$p^{(2)}(x) = \frac{1}{h^2}\left(\delta^2 y_0 + k\delta^3\mu y_0 + \frac{6k^2 - 1}{12}\delta^4 y_0 + \cdots\right),$$

$$p^{(3)}(x) = \frac{1}{h^3}(\delta^3\mu y_0 + k\delta^4 y_0 + \cdots),$$

$$p^{(4)}(x) = \frac{1}{h^4}(\delta^4 y_0 + \cdots) \qquad \text{等等.}$$

13.4 用题 13.3 中的公式，根据表 13.1 的数据来求 $p'(1.10)$，$p^{(2)}(1.10)$和 $p^{(3)}(1.10)$.

解 ☞ 取 $k=0$，在 $x_0=1.10$，我们的公式得到

$$p'(1.10) = 20\left[\frac{0.02411+0.02357}{2} + 0 - \frac{1}{6}\left(\frac{0.00005+0.00004}{2}\right)\right] = 0.4766$$

$$p^{(2)}(1.10) = 400(-0.00054+0) = -0.216$$

$$p^{(2)}(1.10) = 8000(0.000045) = 0.360$$

准确值是 $y'(1.10) = 0.47674$，$y^{(2)}(1.10) = -0.2167$，$y^{(3)}(1.10) = 0.2955$

输入数据精确到五位，但最初三个导数的近似大致分别近似到四位，三位和一位.

13.5 上述问题启示近似微分并不精确. 通过比较函数 $y = e\sin(\pi/e^2)$和近似多项式 $p(x)=0$ 进一步说明这个问题.

解 ☞ 对整数 i，这两个函数在等间距的自变量 $x = ie^2\pi$ 处是配置的. 对很小的数 e，近似是非常精确的，$y(x)-p(x)$永远超不过 e. 但是因为 $y'(x) = \left(\frac{1}{e}\right)\cos(x/e^2)$与 $p'(x)=0$，所以其导数相差很大. 这个例子说明不能预期函数的精确近似就表示其导数的精确近似. 见图 13.1.

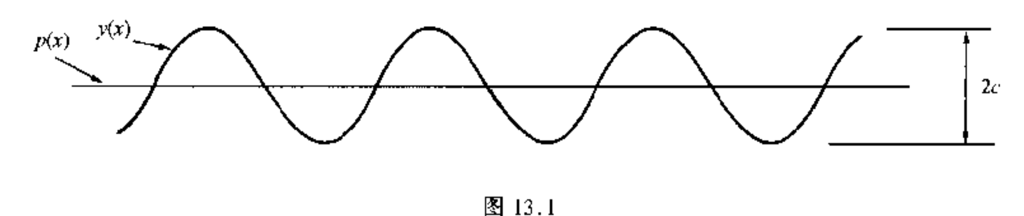


图 13.1

13.6 题 13.1，13.3 和 13.23 启示对 $y'(x_0)$的三种近似仅用一阶差商

$$\frac{y_1-y_0}{h}, \qquad \frac{y_1-y_{-1}}{2h}, \qquad \frac{y_0-y_{-1}}{h}.$$

从几何上讲，它们是图 13.2 中三条直线的斜率，也表示了曲线在 $x=x_0$ 处的切线. 可以认为中间的近似最接近切线的斜率. 可通过计算这三个公式的截断误差证实这一点.

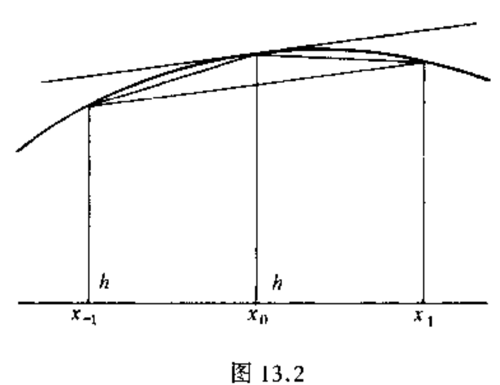


图 13.2

解 ☞ Newton 向前公式，在一阶差商项之后截断，舍去截断误差

$$y(x) - p(x) = \frac{h^2}{2}[k(k-1)y^{(2)}(\xi)],$$

通常取 $x = x_0 + kh$. 这里认为 k 不再限制为整数而是连续变量是有益的. 假设 $y^{(2)}(\xi)$连续，于是得到我们的导数公式的误差(通过连锁法 chain rule). 对 $k=0$,

$$y'(x_0) - p'(x_0) = -\frac{h}{2}y^{(2)}(\xi_0)$$

注意对 $k=0$ 不再包含麻烦的 $y^{(2)}(\xi)$因子的导数. 同样对 Newton 向后公式,

$$y'(x_0) = p'(x_0) - \frac{h}{2}y^{(2)}(\xi_0).$$

应用Stirling公式我们将得到意料之外的好处.即使在我们的近似中保留到二阶差商项,我们发现在 $k=0$ 处它从 $p'(x)$ 消失了.(见题13.3)我们可以认为讨论中的中等近似来自二次多项式近似.于是其截断误差是

$$y(x) - p(x) = \frac{h^3}{6}[(k+1)k(k-1)y^{(3)}(\xi)],$$

可导出

$$y'(x_0) = p'(x_0) = \frac{-h^2}{6}y^{(3)}(\xi).$$

在以上三种计算中,ξ 可能表示三个不同的未知数.但是因为 h 一般很小,出现在最后结果中的 h^2 与出现在别处的 h 相比在"阶量级意义"下,这里的截断误差是最小的.这就证实了几何形象.

13.7 对于表13.1中的数据,应用题13.6中间的公式求 $y'(1.10)$ 的近似.求出这结果的实际误差并与题13.6中的截断误差估计进行比较.

解 ☞ 这一近似实际上是在题13.4中计算 $y'(1.10)\approx 0.4768$ 的第一项.实际误差达到五位,

$$y'(1.10) = 0.4768 = 0.47674 - 0.47680 = -0.00006.$$

在题13.6中得到的估计是 $-h^2y^{(3)}(\xi)/6$,因为 $y^{(3)} = \frac{3}{8}x^{-5/2}$.我们用1代替未知值 ξ 得到 $-h^2y^{(3)}(\xi)/6 \approx -(0.05)^2\left(\frac{1}{6}\right) = -0.00016$ 只是稍微扩大了一点.虽然不是不可能的,这种估计很大方.

13.8 把在题13.3中得到的 $p'(x_0)$ 的公式变换为出现 y_k 值而不是差分形式.

解 ☞ 在这种情形,我们取 $k=0$ 得

$$p'(x_0) = \frac{1}{h}\left[\frac{1}{2}(y_1 - y_{-1}) - \frac{1}{12}(y_2 - 2y_1 + 2y_{-1} - y_{-2})\right]$$
$$= \frac{1}{12}(y_{-2} - 8y_{-1} + 8y_1 + y_2).$$

13.9 估计题13.8中公式的截断误差.

解 ☞ 因为这个公式是基于四次Stirling多项式,

$$y(x) - p(x) = \frac{h^5(k^2-4)(k^2-1)ky^{(5)}(\xi)}{120},$$

如在题13.6中那样求导数,令 $k=0$ 得 $y'(x_0) - p'(x_0) = h^4y^{(5)}(\xi)/30$.

13.10 把题13.9中的估计与在题13.4中的计算结果的实际误差进行比较.

解 ☞ 精确到五位的实际误差是

$$y'(1.10) - p'(1.10) = 0.47674 - 0.47660 = 0.00014.$$

在题13.9中的公式里用 $y^{(5)}(1)$ 代替未知的 $y^{(5)}(\xi)$,稍作扩大后得到

$$\frac{h^4y^{(5)}(\xi)}{30} \approx (0.05)^4\left(\frac{7}{64}\right) = 0.0000007.$$

当然这令人失望,采用高阶差分虽然基本上消除了截断误差但实际误差更大.显然在这一算法中另一种误差源是主要的.这就是 y_i 值的输入误差,并且算法怎样将它们放大.为简单起见将把它归入舍入误差项.

13.11 对公式 $(y_1 - y_{-1})/2h$ 中舍入误差行为进行估计.

解 ☞ 如前,令 Y_1 和 Y_{-1} 是精确的(未知)数据值.于是 $Y_1 = y_1 + e_1$, $Y_{-1} = y_{-1} + e_{-1}$.以 e_1 和 e_{-1} 表示数据误差.差

$$\frac{Y_1 - Y_{-1}}{2h} - \frac{y_1 - y_{-1}}{2h} = \frac{e_1 - e_{-1}}{2h}$$

就是由于输入的不精确而引起的输出误差.如果 e_1 和 e_{-1} 的大小不超过 E,那么输出误差最大是 $\frac{2E}{2h}$,产生最大的舍入误差 E/h.

13.12 应用题 13.11 中的估计来计算题 13.7.

解 ☞ 这里 $h=0.05$, $E=0.00005$, 使得 $E/h=0.00010$. 比算法中的舍入误差可能对第四位稍有影响.

13.13 对题 13.8 中公式中舍入误差影响进行估计.

解 ☞ 确如题 13.11 中那样, 我们求得由于输入的不精确而引起的输出误差是 $(1/12h)(e_{-2}-8e_{-1}+8e_1-e_2)$. 如果 e_k 的大小不超过 E, 这一输出误差最大是 $18E/12h$, 即最大舍入误差 $=(3/2h)E$. 如同题 13.11 中的 $\frac{1}{h}$, 因子 $(3/2h)$ 是放大因子. 对于小的 h, 我们一般把这与高精度联系在一起, 这一因子是大的, 因此输入信息的误差就大大地放大了.

13.14 应用题 13.13 中的估计来计算题 13.4. 然后将我们计算 $y'(1.10)$ 与所关联的各种误差进行比较.

解 ☞ 取 $h=0.05$, $E=0.000005$, $(3/2h)E=0.00015$, 各种误差分组列于表 13.2.

表 13.2

公式	实际误差	估计截断误差	最大舍入误差
$(y_1-y_{-1})/2h$	−0.00006	−0.00016	±0.00010
$(y_{-2}-8y_{-1}+8y_1-y_2)/2h$	0.00014	0.0000007	±0.00015

在第一种情形中舍入误差有所抑制, 而在第二种情形, 舍入误差受到不良影响. 显然, 除了对于极其精确的数据, 这些误差的放大使得不可能产生低的截断误差.

13.15 对以下公式

$$y^{(2)}(x_0) \approx \frac{1}{h^2}\delta^2 y_0 = \frac{1}{h^2}(y_1 - 2y_0 - y_{-1})$$

估计截断误差, 上述公式是题 13.3 中取到第二项商差而得到的.

解 ☞ 这里用不同的方法即用 Taylor 级数求截断误差可能是方便的, 特别地

$$y_1 = y_0 + hy_0' + \frac{1}{2}h^2 y_0^{(2)} + \frac{1}{6}h^3 y_0^{(3)} + \frac{1}{24}h^4 y^{(4)}(\xi_1),$$

$$y_{-1} = y_0 - hy_0' + \frac{1}{2}h^2 y_0^{(2)} - \frac{1}{6}h^3 y_0^{(3)} + \frac{1}{24}h^4 y^{(4)}(\xi_2),$$

把这两项相加并减去 $2y_0$ 就得到

$$\delta^2 y_0 = h^2 y_0^{(2)} + \frac{1}{24}h^4[y^{(4)}(\xi_1) + y^{(4)}(\xi_2)].$$

不幸的是 ξ_1 可能与 ξ_2 不同, 但为了估计截断误差, 假定我们用一个数 $y^{(4)}$ 来代替两个四阶导数, 这种选择仍是不确定的. 为了绝对安全, 我们能在整个区间取 $y^{(4)}=\max|y^{(4)}(x)|$, 这样就得到了截断误差大小的上界, 然而可以想象其它选择也是可以的, 我们现在得到

$$截断误差 = y_0^{(2)} - \frac{1}{h^2}\delta^2 y_0 = -\frac{h^2}{12}y^{(4)}.$$

13.16 用题 13.15 中的估计来计算题 13.4.

解 ☞ 因为更高阶差商项没起作用, 所以在题 13.4 中实际上是通过公式

$$p^{(2)}(1.10) = \frac{\delta^2 y_0}{h^2} = -0.21600$$

来计算 $p^{(2)}(1.10)$, 其结果已与准确值 $y''(1.10)=-0.21670$ 比较. 取 $y^{(4)}(x) = -\frac{15}{16}x^{-\frac{7}{2}} \approx -\frac{15}{16}$, 题 13.15 中的截断误差可能稍微扩大

$$截断误差 \approx \frac{1}{5120} = 0.00020,$$

实际误差是 −0.00070, 再次表明截断误差并非主要的误差源.

13.17 估计公式 $\delta^2 y_0/h^2$ 的截断误差.

解 ☞ 如前进行，我们求得由于输入的不精确而引起的输出误差是$(1/h^2)(e_1-2e_0+e_1)$，其中e_k是输入误差．如果它们的大小都不超过E，则最大输出误差是$(4/h^2)E$．因此，最大的舍入误差$-(4/h^2)E$．

13.18 应用题13.17中的公式计算题13.4，并把我们的近似值与$y^{(2)}(1.10)$之间的实际误差与截断误差和舍入误差估计进行比较．

解 ☞ 如前，取$h=0.05$，　$E=0.000005$，　得到$(4/h^2)E=0.00800$

放大因子$(4/h^2)$有很大影响，我们的结果证实舍入已成为我们在求$y^{(2)}(1.10)$的近似中的主要误差源，而且实际误差和截断误差在大到800单位的舍入误差中仅占了90单位．

实际误差	估计截断误差	最大舍入误差
−0.00070	0.00020	−0.00800

13.19 用题9.7和9.8中的样条求正弦函数导数的近似．

解 ☞ 在题9.7中，我们发现自然样条在端点的二阶导数为零．因为正弦函数本身在端点有同样的二阶导数值，所以在此情形，自然样条是适宜的．先取中心点，我们得到中央样条段S_2的导数是

$$S_2'(x)=\frac{-27\sqrt{3}}{10\pi^3}(2\pi x-\pi^2),$$

它在$x=\frac{\pi}{2}$处精确等于0，显然对称是有益的．在结点$x=\frac{\pi}{3}$处作直接计算可得到S_2'等于0.496这里仅用了在$(0,\pi)$中的三节样条段我们就能下结论，其误差是0.4%．

在题9.8中，我们得到了与正弦函数的一阶导数相匹配的样条，对于中央部分我们求得

$$S_2'(x)=\frac{2\pi\cdot 9\sqrt{3}}{2\pi^3}(2\pi x-\pi^2),$$

它在$x=\frac{\pi}{2}$处也等于0．在$x=\frac{\pi}{3}$处，它等于$(9\sqrt{3}-2\pi)/6\pi$或者0.494．对于二阶导数又出现了预料中的精度降低，在整个中央区间，自然样条的二阶导数$S_2''=-0.948$，而真正的二阶导数是从-0.866到-1．

13.20 如何把Richardson外推法用于数值微分？

解 ☞ 通常近似公式中的误差信息是用于校正的．作为说明，取中心差公式

$$y'(x)=\frac{y(x+h)-y(x-h)}{2h}+T,$$

其中T是截断误差，利用Taylor级数，容易算得

$$T=a_1h^2+a_2h^4+a_3h^6+\cdots.$$

分别对h和$h/2$两次用上式得

$$y'(x)=F(h)+a_1h^2+a_2h^4+\cdots,$$

$$y(x)=F\left(\frac{h}{2}\right)+\frac{a_1h^2}{4}+\frac{a_2h^4}{16}+\cdots,$$

其中$F(h)$和$F\left(\frac{h}{2}\right)$表示近似导数．假定对小的$h$，$a_i$改变不大．消去$a_1$项得

$$y'(x)=\frac{4F(h/2)-F(h)}{3}+b_1h^4+O(h^6).$$

记

$$F_1\left(\frac{h}{2}\right)=\frac{4F(h/2)-F(h)}{3},$$

我们得到了四阶精度的近似微分公式．

现在可重复上述步骤，从

$$y'(x)=F_1\left(\frac{h}{2}\right)+b_1h^4+O(h^6),$$

$$y'(x) = F_1\left(\frac{h}{4}\right) - \frac{b_1 h^4}{16} + O(h^6)$$

开始,消去 b_1 项得到具有六阶精度的近似

$$F_2\left(\frac{h}{2}\right) = \frac{16F_1(h/4) - F_1(h/2)}{15},$$

显然可进一步重复上述过程,这称之为外推到极限.

在外推到极限过程中近似计算的集合一般表示如下:

	F	F_1	F_2	F_3
h	$F(h)$			
$h/2$	$F(h/2)$	$F_1(h/2)$		
$h/4$	$F(h/4)$	$F_1(h/4)$	$F_2(h/4)$	
$h/8$	$F(h/8)$	$F_1(h/8)$	$F_2(h/8)$	$F_3(h/8)$

需要时可加入更多的表值,一般公式是

$$F_m\left(\frac{h}{2^k}\right) = F_{m-1}\left(\frac{h}{2^k}\right) + \frac{F_{m-1}(h/2^k) - F_{m-1}(h/2^{k-1})}{2^{2m} - 1}.$$

不难修改上述过程使得用某些其它方法减小步长,多半是取 $h_i = r^{i-1}h_1$,其中 h_1 是初始步长 h.不用多大的代价就可以处理任意的序列 h_i.有例子说明有时这些变化可能是有益的.

13.21 对函数 $y(x) = -1/x$,用 Richardson 外推法求 $y'(0.05)$.准确值是 400.

解 ☞ 计算结果摘录在表 13.3 中.用的是八位数字计算机.题 13.20 中原有的公式得到 F 这一列(表中所有值缩小 400 倍).对 $h = 0.0001$ 最好的结果在三位小数之内.其后舍入误差占优势.遍观此表其余各列,可见各值几乎可精确到五位.

表 13.3

h	F	F_1	F_2	F_3
0.0128	28.05289			
0.0064	6.66273	−0.46732		
0.0032	1.64515	−0.02737	0.00196	
0.0016	0.41031	−0.00130	0.00043	0.00041
0.0008	0.10250	−0.00010	−0.00002	−0.00002
0.0004	0.02625	0.00084	0.00090	0.00091
0.0002	0.00750	0.00125	0.00127	0.00127
0.0001	0.00500	0.00417	0.00436	0.00441
0.00005	0.01000	0.01166	0.01215	0.01227

表值缩小了 400

补 充 题

13.22 微分 Bessel 公式,得到用直到五阶差分表示的直到 $p^{(5)}(x)$ 的各阶导数.

13.23 用上述问题的结果,根据表 13.1 中的数据去求 p',$p^{(2)}$ 和 $p^{(3)}$ 在 $x = 1.125$ 处的值.

13.24 求在题 13.22 中取 $k = \frac{1}{2}$ 而得到的 $p'(x)$ 的公式的截断误差.估计时取 $\xi = 1$,并与实际误差比较

13.25 求上述题中公式的舍入误差可能的最大值.把实际误差与截断误差和舍入误差的估计值进行比较.

13.26 证明六次 Stirling 公式可得到

$$p'(x_0) = \frac{1}{h}\left(\delta\mu y_0 - \frac{1}{6}\delta^3\mu y_0 + \frac{1}{30}\delta^5\mu y_0\right).$$

证明这个公式的截断误差是 $-h^6y^{(7)}(\xi)/140$.

13.27 把上述公式变换成以下形式

$$p'(x_0)=\frac{1}{60h}(-y_{-3}+9y_{-2}-45y_{-1}+45y_1-9y_2+y_3),$$

并证明最大的舍入误差是 $11E/6h$.

13.28 用 Lagrange 或 Everett 公式(见题 12.11 和 12.12),通过反立方插值,求出表 13.4 中相应于 $y'=0$ 的自变量值,然后通过直接插值求出相应的 y 值.

13.29 不管表 13.4 的首末两行,用 Hermite 公式求和其余数据拟合的三次多项式.这个三次多项式的导数在何处等于零?并与上述问题比较(这里的数据相应于 $y(x)=\sin x$,因此,准确的自变量值是 $\pi/2$.)

13.30 正态分布函数在 $x=1$ 处有一个拐点,如何独立地从以下四位数的数据表中的每一个,精确地确定这个值

表 13.4

x	y	y'
1.4	0.98545	0.16997
1.5	0.99749	0.07074
1.6	0.99957	-0.02920
1.7	0.99166	-0.12884

x	y	x	y
0.50	0.3521	0.98	0.2468
0.75	0.3011	0.99	0.2444
1.00	0.2420	1.00	0.2420
1.25	0.1827	1.01	0.2396
1.50	0.1295	1.02	0.2371

13.31 由题 13.9 和 13.13 我们求得近似等式

$$y'(x_0)\approx\frac{1}{12h}(y_{-2}-8y_{-1}+8y_1-y_2)$$

的合成截断误差和舍入误差具有形式 $Ah^4+3E/2h$,其中 $A=|y^{(5)}(\xi)/30|$.问步长 h 等于多少时,此值最小?对平方根函数计算使结果是五位精度.

13.32 证明公式 $y^{(4)}(x_0)=\delta^4y_0/h^4$ 的截断误差是 $h^2y^{(5)}(\xi)/6$.

13.33 证明题 13.32 中的最大舍入误差是 $16E/h^4$.

第十四章　数值积分

只要注意到在应用分析问题的阐述中是多么频繁地涉及导数，就能意识到数值积分的重要性．人们自然会预期到与导数有关的问题的解将涉及积分．对大多数积分而言是不能用初等函数表示的，数值逼近就成为必要的了．

多项式逼近

多项式逼近作为广泛的一大类积分公式的基础，其主要的思想是假设 $p(x)$ 是 $y(x)$ 的一个逼近，则

$$\int_a^b p(x)\mathrm{d}x \approx \int_a^b y(x)\mathrm{d}x.$$

总体而言，这种逼近是十分成功的．在数值分析中积分是"容易"的一种运算而微分则是"困难"的，而在初等微积分中则或多或少是相反的情况．最知名的例子如下面所列：

1. 在 x_0 与 x_n 之间（在配置点的整个区域），**积分 Newton** 向前公式（n 次）导出几个有用的公式，包括

$$\int_{x_0}^{x_1} p(x)\mathrm{d}x = \frac{h}{2}(y_0 + y_1),$$

$$\int_{x_0}^{x_2} p(x)\mathrm{d}x = \frac{h}{3}(y_0 + 4y_1 + y_2),$$

$$\int_{x_0}^{x_3} p(x)\mathrm{d}x = \frac{3h}{8}(y_0 + 3y_1 + 3y_2 + y_3).$$

   当 $n=1,2$ 及 3．任一个这种公式的截断误差是

$$\int_{x_0}^{x_n} y(x)\mathrm{d}x - \int_{x_0}^{x_n} p(x)\mathrm{d}x.$$

   它可以用不同的方法加以估计，例如，由 Taylor 级数可论证这个误差当 $n=1$ 时近似地为 $-h^3y^{(2)}(\xi)/12$，而当 $n=2$ 时近似地为 $-h^5y^{(4)}(\xi)/90$．

2. **组合公式**是通过将刚才所展示的简单公式反复地加以应用来覆盖一个长区间．它相当于用几个联接的直线段或是抛物线段等等来作逼近，与用单个的高次多项式相比具有简单化的优点．

3. **梯形法则**

$$\int_{x_0}^{x_n} y(x)\mathrm{d}x \approx \frac{1}{2}h(y_0 + 2y_1 + \cdots + 2y_{n-1} + y_n)$$

   是一个基本的，然而是典型的组合公式．当然，它是用直线段联接起来作为对 $y(x)$ 的逼近，它的截断误差近似地为 $-(x_n - x_0)h^2y^{(2)}(\xi)/12$．

4. **Simpson 法则**，$\int_{x_0}^{x_n} y(x)\mathrm{d}x \approx \frac{h}{3}(y_0 + 4y_1 + 2y_2 + 4y_3 + \cdots + 2y_{n-2} + 4y_{n-1} + y_n)$ 也是一种组合公式，但是来自用抛物线段的联接作为对 $y(x)$ 的逼近．这是用来逼近积分的最有力的公式之一．其截断误差约为 $-(x_n - x_0)h^4y^{(4)}(\xi)/180$．

5. **Romberg 方法**是基于梯形法则的截断误差接近正比于 h^2 的事实．将 h 折半并重复应用这个公式，因此误差减少了一个 $\frac{1}{4}$ 因子．将这两个结果进行比较导出对残留误差的一个估计．这个估计可作为一个校正量．Romberg 公式便是将这个简单思想化作系统

精致的改进.

6. **更为复杂的公式**可以通过在比整个配置区域小的区间上,对配置多项式进行积分而得到.例如,带有校正项的 Simpson 法则可以由积分 6 阶的 Stirling 公式而导出,它提供在 $x_{-3}\cdots,x_3$ 点上的配置正好覆盖中心的两个区间 x_{-1}到 x_1,然后将该结果用来展开一个复合公式.这结果为

$$\int_{x_0}^{x_n}y(x)\mathrm{d}x\approx\frac{h}{3}(y_0+4y_1+2y_2+\cdots+y_n)-\frac{h}{90}(\delta^4y_1+\delta^4y_3+\cdots+\delta^4y_{n-1})+\frac{h}{756}(\delta^6y_1+\delta^6y_3+\cdots+\delta^6y_{n-1}),$$

这公式的第一部分便是 Simpson 法则.

7. **Gregory's 公式**取梯形法则的形式再加上校正项.它可以从 Euler-Maclaurin 公式导出.将式中所有的导数表成差分的适当组合而得到:

$$\int_{x_0}^{x_n}y(x)\mathrm{d}x\approx h/2(y_0+2y_1+\cdots+2y_{n-1}+y_n)-\frac{h}{12}(\nabla y_n-\Delta y_0)-\frac{h}{24}(\nabla^2y_n+\Delta^2y_0)-\frac{19h}{720}(\nabla^3y_n-\Delta^3y_0)-\cdots$$

再一次看到其第一部分为梯形法则.Euler-Maclanrin 公式本身可以作为一种近似积分公式.

8. **Taylor 定理**可用来把被积函数展成幂级数,然后逐项积分它有时可导出适宜的积分计算.使用这定理的更精致的方法已被开发出来.

9. 待定系数法可以用来产生具有特殊目的的广泛类型的积分公式.

10. **自适应积分**包含许多已设计出的方法.这些方法处理如下情况:大多数函数在某些区间作精确数值积分要比在其他区间难得多.例如,特殊困难段会迫使在 Simpson 法则中用非常小的 h 值而导致大量非必要的计算.自适应方法只在实在需要的地方才作更细的细分.作此处理的一种系统性的方法将用例子来说明.

误差来源

通常的误差来源依然存在.然而,数据值 $y_0,\cdots,y_n$ 中的输入误差不会被大多数积分公式所放大,所以误差的这种来源不会像在数值微分中那样带来那么多的麻烦.截断误差为

$$\int_a^b[y(x)-p(x)]\mathrm{d}x$$

对于最简单公式以及大多数由相似片段组成的其他公式而言,如今它是主要的误差源.为了估计这个误差已经作了多种多样的努力.一个相关的问题是收敛性.它要问的是,使用的多项式的次数越来越高时,或者使用的数据点之间隔 h_n 越来越小($\lim h_n=0$)时,对于所产生的逼近值序列而言截断误差的极限是否为零.在很多情况下,梯形法则和 Simpson 法则是绝好的例子,其收敛性可以得到证明.然而舍入误差也有强的影响.一个小的步长 h 意味着大量地进行计算因而有许多舍入误差.

这些算法误差最终地掩盖了原本理论上会出现的收敛,并且在实践中发现 h 减少到某个水平之下时会使误差更大而不是更小.截断误差变得可忽略时舍入误差却积累起来,限制了所用方法可能得到的精度.

题 解

14.1 对一个 n 次配置多项式的 Newton 公式进行积分.用上、下限 x_0 及 x_n,它们是配置点的外限,假设等距的节点.

解 ☞ 它包含了积分一个线性函数从 x_0 到 x_1,或是积分一个二次函数从 x_0 到 x_2,等等.参见图

14.1.

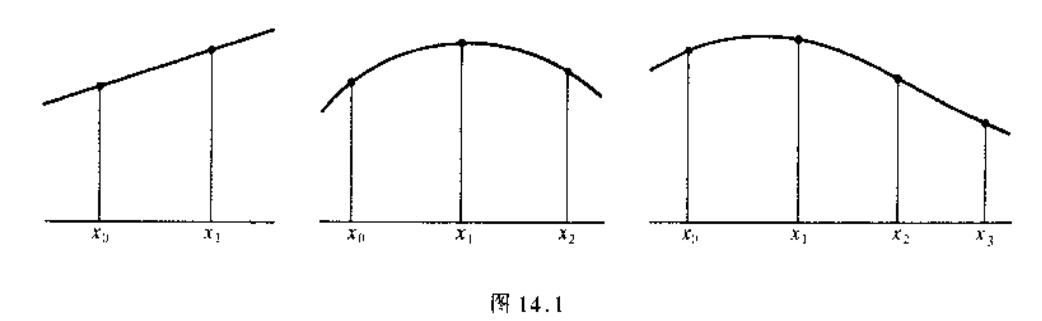


图 14.1

线性函数肯定地导出$\frac{1}{2}h(y_0+y_1)$. 对二次式有

$$p_k = y_0 + k\Delta y_0 + \frac{1}{2}k(k-1)\Delta^2 y_0.$$

由于 $x - x_0 = kh$, 方便的计算产生,

$$\int_{x_0}^{x_2} p(x)\mathrm{d}x = h\int_0^2 p_k \mathrm{d}k = h\left(2y_0 + 2\Delta y_0 + \frac{1}{3}\Delta^2 y_0\right)$$

$$= \frac{h}{3}(y_0 + 4y_1 + y_2).$$

对三次多项式一个类似的计算产生

$$\int_{x_0}^{x_3} p(x)\mathrm{d}x = h\int_0^3 p_k \mathrm{d}k = h\int_0^3 \left[y_0 + k\Delta y_0 + \binom{k}{2}\Delta^2 y_0 + \binom{k}{3}\Delta^3 y_0\right]\mathrm{d}k$$

$$= h\left(3y_0 + \frac{9}{2}\Delta y_0 + \frac{9}{4}\Delta^2 y_0 + \frac{3}{8}\Delta^3 y_0\right) = \frac{3h}{8}(y_0 + 3y_1 + 3y_2 + y_3).$$

高次多项式的结果也可以用同样的形式

$$\int_{x_0}^{x_n} p(x)\mathrm{d}x = Ch(c_0 y_0 + \cdots + c_n y_n)$$

来得到, 对头几个 n 值的 C 及 c_i 在表 14.1 中给出. 这类公式称为 **Cotes 公式**.

表 14.1

n	C	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
1	1/2	1	1							
2	1/3	1	4	1						
3	3/8	1	3	3	1					
4	2/45	7	32	12	32	7				
6	1/140	41	216	27	272	27	216	41		
8	4/14,175	989	5888	−928	10,496	−4540	10,496	−928	5888	989

很少用到较高次的公式, 部分由于有较简单的相同精度的公式可用, 而部分则因为多少有点惊讶的事实是较高次的多项式并不总会改进精度.

14.2　对 $n=1$ 公式估计截断误差

解　对这种简单情况我们可以直接积分公式

$$y(x) - p(x) = \frac{1}{2}(x - x_0)(x - x_1)y^{(2)}(\xi)$$

并应用中值定理如下, 得到精确误差为:

$$\int_{x_0}^{x_1} y(x)\mathrm{d}x - \frac{1}{2}h(y_0 + y_1)$$

$$= \int_{x_0}^{x_1} \frac{1}{2}(x - x_0)(x - x_1)y^{(2)}(\xi)\mathrm{d}x$$

$$= y^{(2)}(\xi)\int_{x_0}^{x_1} \frac{1}{2}(x - x_0)(x - x_1)\mathrm{d}x$$

$$= -\frac{1}{12}h^3y^{(2)}(\xi).$$

其中 $h=x_1-x_0$.由于$(x-x_0)(x-x_1)$在(x_0,x_1)中不改变符号,还有 $y^{(2)}(\xi)$的连续性,应用中值定理是可能的.当 $n>1$ 时一个符号的改变妨碍中值定理的类似应用,而已经设计出许多用来估计截断误差的方法,但大多具有某些缺点.对现在简单的情况 $n=1$ 用 Taylor 级数我们对最古老的方法之一来加以说明.首先我们有

$$\frac{1}{2}h(y_0+y_1)=\frac{1}{2}h\left[y_0+\left(y_0+hy_0'+\frac{1}{2}h^2y_0^{(2)}+\cdots\right)\right]$$

用一个不定积分 $F(x)$,其中 $F'(x)=y(x)$,我们还发现

$$\int_{x_0}^{x_1}y(x)\mathrm{d}x=F(x_1)-F(x_0)=hF'(x_0)+\frac{1}{2}h^2F^{(2)}(x_0)+\frac{1}{6}h^3F^{(3)}(x_0)$$

$$+\cdots=hy_0+\frac{1}{2}h^2y_0'+\frac{1}{6}h^3y_0^{(2)}+\cdots$$

并且相减得

$$\int_{x_0}^{x_1}y(x)\mathrm{d}x-\frac{1}{2}h(y_0+y_1)=-\frac{h^3}{12}y_0^{(2)}+\cdots$$

这将截断误差表示为级数的形式.第一项可以用来作为误差估计.它应该与由 $-(h^3/12)y^{(2)}(\xi)$(其中 $x_0<\xi<x_1$)给出的真正的误差进行比较.

14.3 估计 $n=2$ 公式的截断误差.

解 ☞ 像在前面问题中那样进行,首先我们得到

$$\frac{1}{3}h(y_0+4y_1+y_2)=\frac{1}{3}h\left[y_0+4\left(y_0+hy_0'+\frac{1}{2}h^2y_0^{(2)}\right.\right.$$

$$\left.+\frac{1}{6}h^3y_0^{(3)}+\frac{1}{24}h^4y_0^{(4)}+\cdots\right)+(y_0+2hy_0'+2h^2y_0^{(2)}$$

$$\left.\left.+\frac{4}{3}h^3y_0^{(3)}+\frac{2}{3}h^4y_0^{(4)}+\cdots\right)\right]$$

$$=\frac{1}{3}h\left(6y_0+6hy_0'+4h^2y_0^{(2)}+2h^3y_0^{(3)}+\frac{5}{6}h^4y_0^{(4)}+\cdots\right),$$

该积分本身为

$$\int_{x_0}^{x_2}y(x)\mathrm{d}x=F(x_2)-F(x_0)$$

$$=2hF'(x_0)+\frac{1}{2}(2h)^2F^{(2)}(x_0)+\frac{1}{6}(2h)^3F^{(3)}(x_0)$$

$$+\frac{1}{24}(2h)^4F^{(4)}(x_0)+\frac{1}{120}(2h)^5F^{(5)}(x_0)+\cdots$$

$$=2hy_0+2h^2y_0+\frac{4}{3}h^3y_0^{(2)}+\frac{2}{3}h^4y_0^{(3)}+\frac{4}{15}h^5y_0^{(4)}+\cdots$$

并且相减,

$$\int_{x_0}^{x_2}y(x)\mathrm{d}x-\frac{1}{3}h(y_0+4y_1+y_2)=-\frac{1}{90}h^5y_0^{(4)}+\cdots$$

再一次我们有级数形式的截断误差.第一项将被用作一个近似.还可以被证明误差由 $-(h^5/90)y^{(4)}(\xi)$给出,其中 $x_0<\xi<x_2$.(参看题 14.65)

一个类似的过程应用于其他公式,其结果(只有第一项)列在表 14.2 中

表 14.2

n	截断误差	n	截断误差
1	$-(h^3/12)y^{(2)}$	4	$-(8h^7/945)y^{(6)}$
2	$-(h^5/90)y^{(4)}$	6	$-(9h^9/1400)y^{(8)}$
3	$-(3h^5/80)y^{(4)}$	8	$(2368h^{11}/467\ 775)y^{(10)}$

注意奇次 n 的公式与那些较它小一次的公式相当.(当然,这类公式多包含了一个长度为 h 的区间,然而,这并不证明有什么意义.偶次的公式有优势.)

14.4 导出梯形法则.

解 ☞ 这个古老的公式还是找到它的应用,并且十分简单地说明题 14.1 的公式可以怎样地伸展去覆盖许多区间,梯形法则在逐个区间上用 $n=1$ 的公式直至 x_n.

$$\frac{1}{2}h(y_0+y_1)+\frac{1}{2}h(y_1+y_2)+\frac{1}{2}h(y_2+y_3)+\cdots+\frac{1}{2}h(y_{n-1}+y_n),$$

由它导出公式

$$\int_{x_0}^{x_n}y(x)\mathrm{d}x\approx\frac{1}{2}h(y_0+2y_1+\cdots+2y_{n-1}+y_n),$$

这就是梯形法则.

14.5 将梯形法则用于$\sqrt{x}$在点 1.00 与 1.30 之间的积分.用表 13.1 的数据.与积分的准确值进行比较.

解 ☞ 易见

$$\int_{1.00}^{1.30}\sqrt{x}\,\mathrm{d}x\approx\frac{0.05}{2}[1+2(1.02470+\cdots+1.11803)+1.14017]=0.32147.$$

准确值精确到 5 位是$\frac{2}{3}[(1.3)^{3/2}-1]=0.32149$.

造成真实的误差为 0.00002.

14.6 导出梯形法则截断误差的估计式.

解 ☞ 题 14.2 的结果可以应用于每个区间上,产生的总截断误差约为

$$-\frac{h^3}{12}[y_0^{(2)}+y_1^{(2)}+\cdots+y_{n-1}^{(2)}].$$

假设二阶导数有界,$m<y^{(2)}<M$,方括号中的和将介在 nm 及 nM 之间.同时假设这个导数是连续的,就可以将该和写成 $ny^{(2)}(\xi)$,其中 $x_0<\xi<x_n$.这是因为 $y^{(2)}(\xi)$可以取 m 与M 之间任何值.为方便起见,还可以令积分区间的端点为 $x_0=a$ 及 $x_n=b$,使 $b-a=nh$.把这些综合在一起,我们便有

$$截断误差\approx\frac{(b-a)h^2}{12}y^{(2)}(\xi).$$

14.7 将题 14.6 的估计应用于平方根的积分.

解 ☞ 取 $h=0.05$,$b-a=0.30$ 且 $y^{(2)}(x)=-x^{-3/2}/4$,截断误差≈0.000016,它略小于真正误差 0.00002.然而,舍入到 5 位并且将这个误差加在我们计算所得结果上,就得到 0.32149,这就是准确结果.

14.8 估计在 y_k 值中的不精确性对由梯形法则所得结果的影响.

解 ☞ 以 Y_k 表示真值,如前,我们得到由不精确性 $e_k=Y_k-y_k$ 所造成的误差为$\frac{1}{2}h(e_0+2e_1+\cdots+2e_{n-1}+e_n)$.假如 e_k 的大小不超过 E,则输出误差的界为$\frac{1}{2}h[E+2(n-1)E+E]=(b-a)E$.

14.9 将上面的结果用于题 14.5 的平方根积分.

解 ☞ 我们有$(b-a)E=(0.30)(0.000005)=0.0000015$,故误差的这个来源是可以忽略的.

14.10 导出 Simpson 法则

解 ☞ 它可能是所有积分公式中最流行的一种.它包含了将我们的 $n=2$ 公式逐次应用于区间对直至 x_n,得到的和为

$$\frac{h}{3}(y_0+4y_1+y_2)+\frac{h}{3}(y_2+4y_3+y_4)+\cdots+\frac{h}{3}(y_{n-2}+4y_{n-1}+y_n),$$

它简化为

$$\frac{h}{3}(y_0+4y_1+2y_2+4y_3+\cdots+2y_{n-2}+4y_{n-1}+y_n).$$

这就是 Simpson 法则.它要求 n 为偶数.

14.11 将 Simpson 法则用于题 14.5 的积分.

解 ☞

$$\int_{1.00}^{1.30} \sqrt{x}\,dx = \frac{0.05}{3}[1.0000 + 4(1.02470 + 1.07238 + 1.11803) + 2(1.04881 + 1.09544) + 1.14017] = 0.32149$$

它准确到 5 位.

14.12 估计 Simpson 法则的截断误差.

解 ☞ 题 14.3 的结果可以应用于每一对区间,产生的总截断误差约为

$$-\frac{h^5}{90}(y_0^{(4)} + y_2^{(4)} + \cdots + y_{n-2}^{(4)}).$$

假设 4 次导数连续,在括号中的和允许被写成$(n/2)y^{(4)}(\xi)$,其中 $x_0<\xi<x_n$.(细节几乎与题 14.6 中所述的相同.)由于 $b-a=nh$,

$$截断误差 \approx -\frac{(b-a)h^4}{180}y^{(4)}(\xi).$$

14.13 应用题 14.12 的估计于我们的平方根积分.

解 ☞ 由于 $y^{(4)}(x) = -\frac{15}{16}x^{-7/2}$,截断误差≈0.00000001 它是微不足道的.

14.14 估计数据不精确性对由 Simpson 法则计算结果的影响.

解 ☞ 如在题 14.8 中那样,这个误差为

$$\frac{1}{3}h(e_0 + 4e_1 + 2e_2 + 4e_3 + \cdots + 2e_{n-2} + 4e_{n-1} + e_n).$$

假如数据不精确度 e_k 的大小不超过E,它输出误差界为

$$\frac{1}{3}hE\left[1 + 4\left(\frac{1}{2}n\right) + 2\left(\frac{1}{2}n - 1\right) + 1\right] = (b-a)E,$$

完全像梯形法则那样.将它应用于题 14.11 的平方根积分,我们得到与题 14.9 中所得的一样为 0.0000015,因此又一次说明这种来源的误差可以略去不计.

14.15 比较当小区间分别为 $2h$ 与 h 时应用 Simpson 法则的结果,并获得一个新的截断误差估计.

解 ☞ 假设数据误差可忽略不计,我们比较这二种截断误差,令 E_1 及 E_2 分别表示对小区间分别为 $2h$ 及 h 的误差,于是

$$E_1 = -\frac{(b-a)(2h)^4}{180}y^{(4)}(\xi_1),\quad E_2 = -\frac{(b-a)h^4}{180}y^{(4)}(\xi_2),$$

所以 $E_2=E_1/16$.区间折半时误差减少为原来的$\frac{1}{16}$.现在它可以用来获得 Simpson 法则截断误差的另一种估计.称 I 为积分的精确值,二个 Simpson 逼近值为 A_1 及 A_2,则

$$I = A_1 + E_1 = A_2 + E_2 = A_1 + 16E_2,$$

解出 E_2,与区间 h 相关联的截断误差为$E_2 \approx (A_2 - A_1)/15$.

14.16 用题 14.15 的估计式来校正 Simpson 法则的逼近值.

解 ☞ 这是一个虽初等但十分有用的概念,我们得到

$$I = A_2 + E_2 = A_2 + \frac{A_2 - A_1}{15} = \frac{16A_2 - A_1}{15}.$$

14.17 应用梯形法则,Simpson 法则,及 $n=6$ 的公式从表 14.3 中提供的 7 个值来计算 $\sin x$ 在 0 与 $\pi/2$ 之间的积分.与准确值 1 进行比较.

表 14.3

x	0	$\pi/12$	$2\pi/12$	$3\pi/12$	$4\pi/12$	$5\pi/12$	$\pi/2$
$\sin x$	0.00000	0.25882	0.50000	0.70711	0.86603	0.96593	1.00000

解 ☞ 梯形法则产生的值为 0.99429, Simpson 法则为 1.00003. $n=6$ 的公式导出

$$\frac{\pi}{140(12)}[41(0)+216(25882)+27(5)+272(0.70711)$$
$$+27(0.86603)+216(0.96593)+41(1)]$$
$$=1.000003$$

显见 $n=6$ 的公式对所提供的固定数据完成得最好.

14.18 证明用梯形法则时, 为了得到上题中的积分准确到 5 位的结果, 会要求区间 h 近似于 0.006 弧度. 相反, 表 14.3 中 $h=\pi/12\approx 0.26$

证 ☞ 题 14.6 的截断误差提示我们要有

$$\frac{(b-a)h^2}{12}y^{(2)}(\xi)\leqslant\frac{(\pi/2)h^2}{12}<0.000005$$

即要求 $h<0.006$.

14.19 用 Simpson 法则时, 为了得到题 14.17 中的积分准确到 5 位的结果, 应要求区间 h 为何?

解 ☞ 题 14.12 的截断误差提示

$$\frac{(b-a)h^4}{180}y^{(4)}(\xi)\leqslant\frac{(\pi/2)h^4}{180}<0.000005$$

或近似地 $h<0.15$.

14.20 证明梯形法则与 Simpson 法则是收敛的. 如果我们假设截断为惟一的误差来源, 则在梯形公式中

$$I-A=-\frac{(b-a)h^2}{12}y^{(2)}(\xi).$$

其中 I 为精确积分值, A 为近似值, (此处我们依据的截断误差精确表达式为题 14.2 末尾所提及的.) 若 $\lim h=0$ 并假设 $y^{(2)}$ 有界则 $\lim(I-A)=0$. (这便是收敛的定义.)

证 ☞ 对 Simpson 法则我们有类似的结果

$$I-A=-\frac{(b-a)h^4}{180}y^{(4)}(\xi).$$

若 $h\to 0$ 并假设 $y^{(4)}$ 有界, 则 $\lim(I-A)=0$ 高次公式的多重使用同样也导出收敛性.

14.21 应用 Simpson 法则于积分 $\int_0^{\pi/2}\sin x\,dx$, 在寻找更高精度的过程中将 h 持续地折半.

解 ☞ 以 8 位十进制数进行机器计算, 所产生的结果列于表 14.4 中.

表 14.4

h	近似积分	h	近似积分
$\pi/8$	1.0001344	$\pi/128$	0.99999970
$\pi/16$	1.0000081	$\pi/256$	0.99999955
$\pi/32$	1.0000003	$\pi/512$	0.99999912
$\pi/64$	0.99999983	$\pi/1024$	0.99999870

14.22 题 14.21 的计算指明一种持久的误差来源, 它在 h 变小时不会消失, 当工作继续往下时反而增加. 这是一种什么误差来源?

解 ☞ 对于十分小的区间 h 截断误差是小的, 正如早些时所见, 数据的不精确性对任何 h 来说对 Simpson 法则影响不大. 然而, 小的 h 意味着多的计算, 带来的是大量的舍入误差. 这种误差来源在插值和近似微分时所用的相当简洁的算法中没有成为主要的因素. 这里它变成举足轻重的而且限制了所获得的精度, 即使我们的算法是收敛的(题 14.20)而且数据不精确性的影响也小(我们保存 8 位小数). 这个问题强调了继续寻找更加简洁算法的重要性.

14.23 将题 14.15 及 14.16 中的思想推广到近似积分的 Romberg 方法.

解 ☞ 假定近似公式的误差与 h^n 成正比．于是以 $2h$ 和 h * 为区间长的该公式的两次应用时误差分别为

$$E_1 \approx C(2h)^n, \qquad E_2 \approx Ch^n,$$

得出 $E_2 \approx E_1/2^n$．像前面那样用 $I = A_1 + E_1 = A_2 + E_2$，我们应得新的近似值

$$I \approx A_2 + \frac{A_2 - A_1}{2^n - 1} = \frac{2^n A_2 - A_1}{2^n - 1}.$$

对 $n = 4$ 这就是题 14.16，对 $n = 2$ 它应用于梯形法则，它的截断误差与 h^2 成正比．不难证明当 $n = 4$ 时我们最后的那个公式与 Simpson 法则相吻合，而 $n = 4$ 时它与 $n = 4$ 的 Cotes 公式相吻合．可以证明在这个公式中误差与 h^{n+2} 成正比，而这一点提示了一种递推计算．持续地将 h 折半应用梯形法则若干次．记这些结果为 $A_1, A_2, A_3, \cdots$．将上面我们的公式以 $n = 2$ 应用于相继的 A_i 对，令其结果为 $B_1, B_2, B_3, \cdots$．由于现在的误差正比于 h^4，我们可以应用这个公式于 B_i，取 $n = 4$．这些结果可记为 $C_1, C_2, C_3, \cdots$．按这个模式继续下去，可获一组结果：

$$\begin{array}{ccccc} A_1 & A_2 & A_3 & A_4 & \cdots \\ & B_1 & B_2 & B_3 & \cdots \\ & & C_1 & C_2 & \cdots \\ & & & D_1 & \cdots \end{array}$$

计算持续到数组的右下方成员在所要求的容许量之内．

14.24 应用 Romberg 方法于题 14.21 的积分．

解 ☞ 不同的结果列出如下：

所用的点	4	8	16	32
梯形结果	0.987116	0.996785	0.999196	0.999799
		1.000008	1.000000	1.000000
			1.000000	1.000000
				1.000000

收敛于准确值 1 是明显的．

14.25 在一个小于整个配置区域的区间上积分一个多项式．由此可能得到更精确的积分公式．在二个中心区间上积分 Stirling 公式．

解 ☞ 直到包括 6 阶差分的 Stirling 公式为

$$p_k = y_0 + k\mu\delta y_0 + \frac{1}{2}k^2\delta^2 y_0 + \frac{k(k^2-1)}{6}\mu\delta^3 y_0 + \frac{k^2(k^2-1)}{24}\delta^4 y_0 + \frac{k(k^2-1)(k^2-4)}{120}\mu\delta^5 y_0 + \frac{k^2(k^2-1)(k^2-4)}{720}\delta^6 y_0.$$

由于 $x = x_0 + kh$ 且 $dx = h\,dk$，积分得出

$$\int_{x_0-h}^{x_0+h} p(x)dx = h\int_{-1}^{1} p_k dk = h\left(2y_0 + \frac{1}{3}\delta^2 y_0 - \frac{1}{90}\delta^4 y_0 + \frac{1}{756}\delta^6 y_0\right).$$

通过增加多项式的次数来获得更多项显然是有用的．在二阶差分处截断带给我们的仍是 Simpson 法则的开始组合，其形式为 $(h/3)(y_{-1} + 4y_0 + y_1)$．在这种情况下积分伸展到整个配置区域上，就像在题 14.1 中那样．带上个 4 阶差分项我们只在半个配置区域上进行积分(图 14.2)．

当使用更多的差分时 $y(x)$ 及 $p(x)$ 要在附加的点上进行配置，而积分只伸展到中央的二个区间上．由于这些是 Stirling 公式有最小截断误差的区间(题 12.64)，可以预料以这种方法所得的积分公式更为精确．然而，这额外的精度是要付出代价的；在这类公式的应用中要求在积分区间外的 y_k 值．

---

* 译注：原文为 h 和 $2h$．

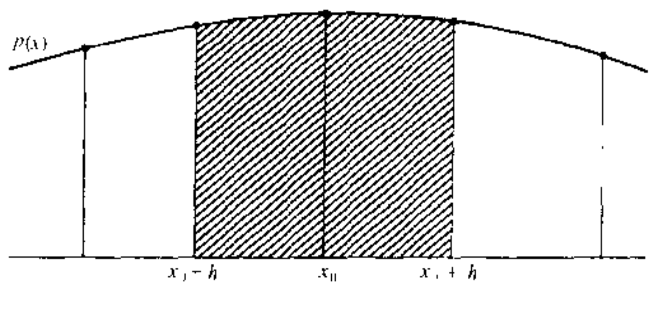


图 14.2

该公式的截断误差可以由题 14.6 中的 Taylor 级数法估得，并证明它近似地为 $-\frac{23h^9}{113400}y_0^{(8)}+\cdots$.

14.26 用题 14.25 中的结果来推出带校正项的 Simpson 法则.

解 ☞ 当 n 为偶数时，我们分别以 $x_1, x_3, \cdots, x_{n-1}$ 为中心点应用上题结果 $n/2$. 得到结果：

$$\int_{x_0}^{x_n} p(x)\mathrm{d}x = \frac{h}{3}(y_0+4y_1+2y_2+\cdots+4y_{n-1}+y_n)$$

$$-\frac{h}{90}(\delta^4 y_1+\delta^4 y_3+\cdots+\delta^4 y_{n-1})$$

$$+\frac{h}{756}(\delta^6 y_1+\delta^6 y_3+\cdots+\delta^6 y_{n-1}).$$

假如需要的话它可以扩展到更高阶差分.

这结果的截断误差约为上题结果的 $n/2$ 倍，因而可以写成 $-\frac{23(x_n-x_0)h^8}{226\,800}y_0^{(8)}+\cdots$.

14.27 发展自适应积分思想.

解 ☞ 自适应的基本的思想是将积分区间的每个部分作细分，让小区间小得足以使其所产生的误差在全局误差中恰如其分. 存在许多途径可完成这件事. 假定全局允许的误差为 E. 选择一个积分公式并将它用于这个区间. 应用一个误差估计式. 若估得它比 E 小，就算完成了. 假如不是，将该公式应用于区间的左半部，如果新的误差估计小于 $E/2$，我们就在那半个区间完成了. 假如还不行，区间再对半分然后进行下去. 最后达到一个长度为 $(b-a)/2^k$ 的区间（(a,b) 为最初的区间），在那里所用公式产生一个可接受的结果，误差小于 $E/2^k$. 接着，再从这个可接受区间的右边界处重新开始这个过程.

作为基本的积分公式，可选 Simpson 法则

$$A_2 = h/3(y_0+4y_1+2y_2+4y_3+y_4)$$

为度量误差，用双倍区间法则

$$A_1 = \frac{2h}{3}(y_0+4y_2+y_4)$$

是方便的. 由题 14.15 于是可估计误差为 $(A_2-A_1)/15$. 因此，每当 $A_2-A_1 \leqslant 15E/2^k$，近似值 A_2 则是可接受的，并且将其累加到从它左边来的可接受结果之和中. 显然，这过程进行到 (a,b) 被可接受的片段全部覆盖为止.

14.28 将上题中的自适性积分法应用于积分

$$\int_0^8 x^5 \mathrm{d}x$$

解 ☞ 以不同的容许量和稍加变化的上限计算几次. 下面扼要的输出是有代表性的. 特别要注意 k 值，它从 1 开始（这里没有打印出来）升到 7. 进一步增加上限 k 将猛增.

r	$r^6/6$	计算值	k
2	10.667	10.667	4
4	682.667	682.667	5
6	7 776.000	7 775.99	6
8	43 690.67	43 690.58	7

14.29 应用自适应积分于反正弦积分

$$\int_0^1 \frac{dx}{\sqrt{1-x^2}}.$$

解 ☞ 在上限处的无穷性间断造成了麻烦,它提示在靠近这个端点处步长将减到非常小,就像上题中那样. k 值在计算的过程中稳步地上升,这个增到 15 时有以下结果:

$$上限 = 0.9999,$$
$$积分 = 1.5573.$$

在这一点上准确的反正弦值为 1.5575.

14.30 导出 Gregory 公式.

解 ☞ 这是梯形法则带有校正项的一种形式,可以用多种方法导出,一种方法是从下面形式的 Euler-Maclaurin 公式开始

$$\int_{x_0}^{x_n} y(x)dx = \frac{h}{2}(y_0 + 2y_1 + \cdots + 2y_{n-1} + y_n)$$
$$- \frac{h^2}{12}(y'_n - y'_0) + \frac{h^4}{720}(y_n^{(3)} - y_0^{(3)})$$
$$- \frac{h^6}{30\,240}(y_n^{(5)} - y_0^{(5)}).$$

假如需要的话取更多的项是有用的.现将 x_n 处的导数以后差形式来表示,而在 x_0 处的导数以前差表示(题 13.1),

$$hy'_0 = \left(\Delta - \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 - \frac{1}{4}\Delta^4 + \frac{1}{5}\Delta^5 - \cdots\right)y_0,$$
$$hy'_n = \left(\nabla + \frac{1}{2}\nabla^2 + \frac{1}{3}\nabla^3 + \frac{1}{4}\nabla^4 + \frac{1}{5}\nabla^5 + \cdots\right)y_n,$$
$$h^3y_0^{(3)} = \left(\Delta^3 - \frac{3}{2}\Delta^4 + \frac{7}{4}\Delta^5 - \cdots\right)y_0,$$
$$h^3y_n^{(3)} = \left(\nabla^3 + \frac{3}{2}\nabla^4 + \frac{7}{4}\nabla^5 + \cdots\right)y_n,$$
$$h^5y_0^{(5)} = (\Delta^5 - \cdots)y_0,$$
$$h^5y_n^{(5)} = (\nabla^5 + \cdots)y_n.$$

将这些表达式代入的结果为

$$\int_{x_0}^{x_n} p(x)dx = \frac{h}{2}(y_0 + 2y_1 + \cdots + 2y_{n-1} + y_n)$$
$$- \frac{h}{12}(\nabla y_n - \Delta y_0) - \frac{h}{24}(\nabla^2 y_n + \Delta^2 y_0)$$
$$- \frac{19h}{720}(\nabla^3 y_n - \Delta^3 y_0) - \frac{3h}{160}(\nabla^4 y_n + \Delta^4 y_0)$$
$$- \frac{863h}{60\,480}(\nabla^5 y_n - \Delta^5 y_0).$$

同样如果需要的话可以计算更多的项.这就是 Gregory 公式,它并不要求积分区间以外的 y_k 值.

14.31 应用 Taylor 定理计算误差函数积分

$$H(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt,$$

对 $x = 0.5$ 及 $x = 1$ 准确到 4 位小数.

解　级数 $e^{-t^2}=1-t^2+\frac{t^4}{2}-\frac{t^6}{6}+\frac{t^8}{24}-\frac{t^{10}}{120}+\cdots$导出

$$H(x)=\frac{2}{\sqrt{\pi}}\left(x-\frac{x^3}{3}+\frac{x^5}{10}-\frac{x^7}{42}+\frac{x^9}{216}-\frac{x^{11}}{1320}+\cdots\right),$$

对 $x=5$,它的结果为 0.5205,而对 $x=1$ 我们得到 0.8427.这个级数的特征保障了由截断造成的误差不会超过所用的最后一项,所以可以对我们的结果有信心.这里级数方法完成得很好.但是显而易见,如果要求更多位数的小数或是使用更大的 x 上限,那么就要取多得多的级数项.在这种情况下如下题那样处理通常更为方便.

14.32　当 $x=0(0.1)4$ 时,对误差函数积分进行 6 位小数的列表.

$$H(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt.$$

解　我们采用在 NBS-AMS41 中准备这个函数的 15 位表时曾用过的方法.所需的导数为

$$H'(x)=\frac{2}{\sqrt{\pi}}e^{-x^2},\quad H^{(2)}(x)=-2xH'(x),$$

$$H^{(3)}(x)=-2xH^{(2)}(x)-2H'(x).$$

而一般地有

$$H^{(n)}(x)=-2xH^{(n-1)}(x)-2(n-2)H^{(n-2)}(x).$$

Taylor 级数可以写成

$$H(x+h)=H(x)+hH'(x)+\cdots+\frac{h^n}{n!}H^{(n)}(x)+R,$$

这里余项通常指 $R=h^{n+1}H^{(n+1)}(\xi)/(n+1)!$.注意到若 M 表示偶次幂项的和而 N 表示奇次项的和,则

$$H(x+h)=M+N,\qquad H(x-h)=M-N.$$

对于精确到 6 位,我们用影响到第 8 位的 Taylor 级数,因为随着计算量的增多将使得舍入误差可能有实质性的增长.取 $H(0)=0$,计算从

$$H(0.1)=\frac{2}{\sqrt{\pi}}(0.1)-\frac{2}{3\sqrt{\pi}}(0.1)^3+\frac{1}{5\sqrt{\pi}}(0.1)^5=0.11246291$$

开始,只有奇次幂起作用.接着我们令 $x=0.1$ 便得到

$$H'(0.1)=\frac{2}{\sqrt{\pi}}e^{-.01}=1.1171516,$$

$$H^{(2)}(0.1)=-0.2H'(0.1)=-.22343032,$$

$$H^{(3)}(0.1)=-0.2H^{(2)}(0.1)-2H'(0.1)=-2.1896171,$$

$$H^{(4)}(0.1)=-0.2H^{(3)}(0.1)-4H^{(2)}(0.1)=1.3316447,$$

$$H^{(5)}(0.1)=-0.2H^{(4)}(0.1)-6H^{(3)}(0.1)=12.871374,$$

$$H^{(6)}(0.1)=-0.2H^{(5)}(0.1)-8H^{(4)}(0.1)=-13.2274320.$$

导致

$$M=0.11246291-0.00111715+0.00000555-0.00000002=0.11135129,$$

$$N=0.11171515-0.00036494+0.00000107=0.11135129.$$

由于 $H(x-h)=M-N$,我们再次发现 $H(0)=0$,它充当对计算准确性的一次检验,再一次我们得到

$$H(0.2)=H(x+h)=M+N=0.22270258$$

现在重复这个过程来得到对 $H(0.1)$的一个检验及对 $H(0.3)$的一个预测,用这种方法进行下去最终达到 $H(4)$.最后二位小数可以舍去.在表 14.5 中给出 $x=0(0.5)4$ 的 6 位准确值.在 NBS-AMS41 中计算进行到 25 位,然后舍入到 15 位.对小的 x 值表要作大量的加密.

表 14.5

x	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
$H(x)$	0.520500	0.842701	0.966105	0.995322	0.999593	0.999978	0.999999	1.000000

14.33 说明导出近似积分公式的待定系数法，把它用于指导 Simpson 法则为例.

解 ☞ 在这个方法中我们直接针对一个预先选定类型的公式. 对于 Simpson 公式选择

$$\int_{-h}^{h} y(x)\mathrm{d}x = h(c_{-1}y_{-1} + c_0y_0 + c_1y_1)$$

是方便的. 系数 c_k 的选择可以用多种方法进行，但是对 Simpson 法则来说，选择是建立在这样的基础上：所获公式当 $y(x)$ 为 x 的头三次幂中任一个时应是精确的. 轮流取 $y(x)=1, x$ 及 x^2，我们得到条件

$$2 = c_{-1} + c_0 + c_1,\quad 0 = -c_{-1} + c_1,\quad \frac{2}{3} = c_{-1} + c_1.$$

其结果为 $c_{-1} = c_1 = \frac{1}{3}$，$c_0 = \frac{4}{3}$，使得

$$\int_{-h}^{h} y(x)\mathrm{d}x = \frac{h}{3}(y_{-1} + 4y_0 + y_1).$$

将这个结果应用于 x_0 与 x_n 之间相继的区间对上，再一次产生 Simpson 法则.

作为额外的收获，这个结果还证明对 $y(x)=x^3$ 也是精确的，这一点从对称性中是容易看到的. 加上这一点意味着对任何不大于 3 次的多项式公式还是精确的. 对于更高次多项式存在误差项.

14.34 应用待定系数法导出型如

$$\int_0^h y(x)\mathrm{d}x = h(a_0y_0 + a_1y_1) + h^2(b_0y_0' + b_1y_1')$$

的公式.

解 ☞ 有 4 个可用系数，我们试着使之当 $y(x)=1, x, x^2$ 及 x^3 时为精确的，它导出的 4 个条件为

$$\begin{aligned}1 &= a_0 + a_1,\\ \frac{1}{2} &= a_1 + b_0 + b_1,\\ \frac{1}{3} &= a_1 + 2b_1,\\ \frac{1}{4} &= a_1 + 3b_1.\end{aligned}$$

由此得出 $a_0 = a_1 = \frac{1}{2}$，$b_0 = -b_1 = \frac{1}{12}$，其结果是公式

$$\int_0^h y(x)\mathrm{d}x = \frac{h}{2}(y_0 + y_1) + \frac{h^2}{12}(y_0' - y_1').$$

这是对 Euler-Maclaurin 公式第一项的复制. 待定系数法可以产生许多不同的公式. 正如在刚才所提供的例子中，稍加计划并利用对称性常常可以简化最终决定系数的方程组.

补 充 题

14.35 对一个 4 次配置多项式积分 Newton 公式，并以此验证表 14.1 的 $n=4$ 行.

14.36 验证表 14.1 的 $n=6$ 行.

14.37 利用 Taylor 级数方法来得到对 $n=3$ 公式的截断误差估计，如表 14.2 所列的.

14.38 利用 Taylor 级数方法验证 $n=4$ 公式的截断误差估计.

14.39 应用不同的公式对下面提供的有限数据去近似 $y(x)$ 的积分.

x	1.0	1.2	1.4	1.6	1.8	2.0
$y(x)$	1.0000	0.8333	0.7143	0.6250	0.5556	0.5000

使用带有校正项的梯形法则.对你的结果你有多大的自信心?它是否准确到4位?(参看下一题)

14.40 题14.39的数据实际是属于函数 $y(x)=1/x$.因此,准确积分到4位,ln2 = 0.6931.是否任意的近似公式都有相同的结果?

14.41 对梯形公式使用截断误差估计来预测 $y(x)$ 值,应挑选得多密(什么样的区间长 h),才使得用梯形公式本身于积分 $\int_1^2 dx/x$ 时能达到4位准确的结果.

14.42 假设题14.39的数据扩大到包括这些新的数对:

x	1.1	1.3	1.5	1.7	1.9
$y(x)$	0.9091	0.7692	0.6667	0.5882	0.5263

对所提供的全部数据再一次应用梯形法则.将这个结果当作 A_2,题14.39中相应的结果当作 A_1,并且用题14.23中的公式来获得 I 的另一个近似.它是否准确到4位?

14.43 应用带有校正项的梯形法则于至今所提供的属于 $y(x)=1/x$ 的全部数据.

14.44 应用Simpson法则于题14.39中的数据.是否像在题14.26中那样需要校正项?如果是,就应用它们.

14.45 对Simpson法则用截断误差估计来预测,对这个法则来说需要多少个 $y(x)$ 的值(或者区间 h 应如何地小),才能使得ln2有准确到4位的结果.

14.46 用梯形法则要得到ln2准确到8位,应要求区间 h 为多大?用Simpson法则呢?

14.47 应用Euler-Maclaurin公式(题14.30)直到5次导数项,估计ln2到8位小数.准确值为0.69314718(试用 $h=0.1$)

14.48 从下面的数据尽你的所能来估计 $\int_0^2 y(x)dx$.

x	0	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2
$y(x)$	1.000	1.284	1.649	2.117	2.718	3.490	4.482	5.755	7.389

你对自己的结果有多大自信心?你相信它们准确到3位吗?

14.49 题14.48的数据是取自于指数函数 $y(x)=e^x$.因此准确到3位的积分为 $\int_0^2 e^x dx = e^2-1=6.389$.是否从我们的任一公式均能获得这个结果?

14.50 从下面的数据,尽你的所能来估计 $\int_1^5 y(x)dx$.

x	1	1.5	2	2.5	3	3.5	4	4.5	5
$y(x)$	0	0.41	0.69	0.92	1.10	1.25	1.39	1.50	1.61

你对自己的结果有多大自信心?

14.51 题14.50的数据对应于 $y(x)=\log x$.因此,准确到2位的积分值为 $\int_1^5 \log x dx = 5\log 5-4=4.05$.是否从我们的任一公式均能获得这个结果?

14.52 以自适应积分计算 $\int_0^1 \frac{dx}{1+x^2}$,准确到7位.准确值为 $\pi/4$,或者算到7位为0.7853982.

14.53 计算 $\int_0^{\pi/2}\sqrt{1-\frac{1}{4}\sin^2 t}\,dt$ 到4位小数.它称为椭圆积分.它的准确值为1.4675.用自适应积分.

14.54 证明若取4位,则 $\int_0^{\pi/2}\sqrt{1-\frac{1}{2}\sin^2 t}\,dt = 1.3506$.

14.55 使用自适应积分来证明

$$\int_0^{\pi/2}\frac{dx}{\sin^2 x+\frac{1}{4}\cos^2 x}=3.1415927.$$

精确值为 π.

14.56 如在题 14.31 中那样应用 Taylor 级数对 $x=0(0.1)1$ 计算正弦积分

$$\mathrm{Si}(x)=\int_0^x\frac{\sin t}{t}dt$$

到 5 位小数.这里不需要用在题 14.32 中的改进过程.[最终结果应是 $\mathrm{Si}(1)=0.94608$.]

14.57 像在题 14.32 中那样应用 Taylor 级数方法对 $x=0(0.5)15$ 计算正弦积分到 5 位小数.最终结果应为 $\mathrm{Si}(15)=1.61819$.

14.58 应用 Taylor 级数方法计算 $\int_0^1\sqrt{x}\sin x\,dx$ 到 8 位小数.

14.59 应用 Taylor 级数方法计算 $\int_0^1\left(1/\sqrt{1+x^4}\right)dx$ 到 4 位小数.

14.60 计算椭圆 $x^2+y^2/4=1$ 的总弧长到 6 位小数.

14.61 通过将表 14.1 的 $n=6$ 公式加上 $(h/140)\delta^6 y_3$ 项导出 Weddle 法则,

$$\int_{x_0}^{x_6}y(x)dx=\frac{3h}{10}(y_0+5y_1+y_2+6y_5+y_4+5y_5+y_6).$$

14.62 使用待定系数法导出一个形如

$$\int_{-h}^{h}y(x)dx=h(a_{-1}y_{-1}+a_0y_0+a_1y_1)+h^2(b_{-1}y'_{-1}+b_0y'_0+b_1y'_1)$$

的公式,它对次数尽可能高的多项式来说是精确的.

14.63 使用待定系数法导出公式

$$\int_0^h y(x)dx=\frac{h}{2}(y_0+y_1)-\frac{h_3}{24}(y_0^{(2)}+y_1^{(2)}).$$

证明它对到 3 次为止的多项式是精确的.

14.64 使用待定系数法导出

$$\int_0^h y(x)dx=\frac{h}{2}(y_0+y_1)+\frac{h^2}{10}(y'_0-y'_1)+\frac{h^3}{120}(y_0^{(2)}+y_1^{(2)}),$$

并证明它对到 5 次为止的多项式是精确的.

14.65 以下面的方法为我们的 $n=2$ 公式的截断误差导出一个精确表达式.令

$$F(h)=\int_{-h}^{h}y(x)dx-\frac{h}{3}[y(-h)+4y(0)+y(h)],$$

用"在积分号下微分"的定理

$$\frac{d}{dh}\int_{a(h)}^{b(h)}y(x,h)dx=\int_a^b\frac{2y}{2h}dx+y(b,h)b'(h)-y(a,h)a'(h).$$

对 h 微分 3 次得到

$$F^{(3)}(h)=-h/3[y^{(3)}(h)-y^{(3)}(-h)].$$

注意 $F'(0)=F^{(2)}(0)=F^{(3)}(0)=0$.假设 $y^{(4)}(x)$ 连续,现在由中值定理得到

$$F^{(3)}(h)=-2/3h^2y^{(4)}(\theta h),$$

其中 θ 依赖于 h 并落在 -1 与 1 之间,现在我们反向进行并通过积分重获 $F(h)$.为方便起见,将 h 换成 t(使 θ 为 t 的函数).证明

$$F(h)=-1/3\int_0^h(h-t)^2t^2y^{(4)}(\theta t)dt$$

通过对 h 微分 3 次来重现上面的 $F^{(3)}(h)$.由于这个公式也使得 $F(0)=F'(0)=F^{(2)}(0)$,所以这就是最初的 $F(h)$.接着应用中值定理

$$\int_a^b f(t)g(t)dt=g(\xi)\int_a^b f(t)dt,$$

具有 $a<\xi<b$,它对假定在 a 与 b 之间符号不变的连续函数是成立的.对于这里的 $f(t)=-t^2(h-t)^2/3$,这些条件确实都成立.其结果为

$$F(h)=y^{(4)}(\xi)\int_0^h f(t)dt=-\frac{h^5}{90}y^{(4)}(\xi),$$

这是题 14.3 中提到的结果.在这证明的开头几步我们从 $F(h)$ 一直推进到它的三阶导数并且再反过来,其目的就是为了得到一个能对它使用中值定理的 $F(h)$ 的表达式.[也就是在积分区间里 $f(t)$ 不变符号.]这往往是为获得一个如刚才所得到的那一类截断误差公式时的中心难点.

14.66 将题 14.65 的论证加以修改以求得在题 14.2 末尾所给出的公式,$n=1$ 时的公式为

$$截断误差 = -\frac{h^3}{12}y^{(2)}(\xi).$$

14.67 计算 $\int_0^1 e^{-x^3}dx$ 准确到 6 位.

第十五章　Gauss 积分

Gauss 公式的特性

Gauss 积分背后的主要思想是在选择公式

$$\int_a^b y(x)\mathrm{d}x \approx \sum_{i=1}^n A_i y(x_i)$$

时可以明智地不去指定自变量 x_i 为等距的.上章中所有公式均假设为等距的,假如 $y(x_i)$值是从实验所获得的,这或许是可实现的.然而,许多积分涉及常用解析函数,它们可以在任何自变量点处进行计算而且达到一个高的精度.在这种情况下可提出如下有益的要求:怎样一并选择 x_i 与A_i 以带来最大的精确度.事实证明讨论更一般些的公式

$$\int_a^b w(x)y(x)\mathrm{d}x \approx \sum_{i=1}^n A_i y(x_i)$$

是方便的,其中 $w(x)$是一个以后要加以规定的权函数.当 $w(x)=1$ 时便回到原先较简单的公式.

对这类 Gauss 型公式的一种处理是要求当 $y(x)$为幂函数 $1,x,x^2,\cdots,x^{2n-1}$之一时,公式为完全精确的.这便对 $2n$ 个数 x_i 及 A_i 提供了 $2n$ 个条件.事实上

$$A_i = \int_a^b w(x)L_i(x)\mathrm{d}x$$

其中 $L_i(x)$为第八章中引进的 Lagrange 乘算子函数.自变量 $x_1,\cdots,x_n$ 为 n 次多项式 $p_n(x)$的零点,$p_n(x)$属于具有正交性的函数族.

$$\int_a^b w(x)p_n(x)p_m(x)\mathrm{d}x = 0 \qquad 当\ m \neq n.$$

这些多项式依赖于 $w(x)$.因此权函数影响 A_i 及 x_i 但在 Gauss 型积分中不明显.

关于密切多项式的 Hermite 公式提供了对 Gauss 型积分的另一种处理.对密切多项式积分导出

$$\int_a^b w(x)y(x)\mathrm{d}x \approx \sum_{i=1}^n [A_i y(x_i) + B_i y'(x_i)].$$

然而,将自变量 x_i 选择为正交族中一员的零点时,会使所有 $B_i=0$.公式就还原成前面给出的类型.它表明了(我们也已证实),对这些非等距的自变量用一个简单的配置多项式也会导出同样的结果.

因此在 Gauss 型积分中正交多项式扮演了一个中心角色.对它们主要性质的研究成为本章的一个实质部分.

Gauss 型公式的截断误差为

$$\int_a^b w(x)y(x)\mathrm{d}x - \sum_{i=1}^n A_i y(x_i) = \frac{y^{(2n)}(\xi)}{(2n)!}\int_a^b w(x)[\pi(x)]^2\mathrm{d}x$$

其中 $\pi(x)=(x-x_1)\cdots(x-x_n)$.由于它与 $y(x)$的 $2n$ 次导数成正比,故这类公式对所有不大于 $2n-1$ 次的多项式是精确的.在上章的公式中在这个位置上出现的是 $y^{(n)}(\xi)$,从某种意义上来讲我们现在的公式二倍精确于那些建立在等距自变量上的公式.

特殊类型的 Gauss 公式

特殊类型 Gauss 公式可以从以不同的方法来选取 $w(x)$与积分上下限而得到.必要时也可以希望加点限制,诸如预先指定某些 x_i.许多特殊类型都已被提出.

1. **Gauss-Legendre 公式**出现于 $w(x)=1$. 这就是典型的 Gauss 方法,我们要比讨论其他类型更细致地讨论它. 习惯地将区间 (a,b) 规范化为 $(-1,1)$. 于是正交多项式就是 Legendre 多项式

$$P_n(x)=\frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n,$$

取 $P_0(x)=1$. x_i 为这些多项式的零点而那些系数为

$$A_i=\frac{2(1-x_i^2)}{n^2[P_{n-1}(x_i)]^2}.$$

x_i 与 A_i 的表可用于直接地代入到 Gauss-Legendre 公式

$$\int_a^b y(x)dx\approx\sum_{i=1}^n A_i y(x_i).$$

Legendre 多项式的各种性质在指导这些结果时是需要的,包括下面这些:

$$\int_{-1}^1 x^k P_n(x)dx=0\qquad 当\ k=0,\cdots,n-1;$$

$$\int_{-1}^1 x^n P_n(x)dx=\frac{2^{n+1}(n!)^2}{(2n+1)!};$$

$$\int_{-1}^1[P_n(x)]^2dx=\frac{2}{2n+1};$$

$$\int_{-1}^1 P_m(x)P_n(x)dx=0,\qquad 当\ m\neq n;$$

$P_n(x)$ 在 $(-1,1)$ 中有 n 个实零点 L;

$$(n+1)P_{n+1}(x)=(2n+1)xP_n(x)-nP_{n-1}(x);$$

$$(t-x)\sum_{i=0}^n(2i+1)P_i(x)P_i(t)=(n+1)[P_{n+1}(t)P_n(x)-P_n(t)P_{n+1}(x)];$$

$$\int_{-1}^1\frac{P_n(x)}{x-x_k}dx=\frac{-2}{(n+1)P_{n+1}(x_k)};$$

$$(1-x^2)P_n'(x)+nxP_n(x)=nP_{n-1}(x).$$

Gauss-Legendre 公式截断误差的 Lanczos 估计取下面的形式

$$E\approx\frac{1}{2n+1}\Big[y(1)+y(-1)-I-\sum_{i=1}^n A_i x_i y'(x_i)\Big],$$

其中 I 是由 Gauss n-点公式所获得的近似积分. 注意到 $\sum$ 项含有对函数 $xy'(x)$ 应用同样的公式. 这个误差估计对光滑函数而言看来十分精确.

2. **Gauss-Laguerre 公式**取形式为

$$\int_0^\infty e^{-x}y(x)dx\approx\sum_{i=1}^n A_i y(x_i),$$

自变量 x_i 为 n 次 Laguerre 多项式

$$L_n(x)=e^x\frac{d^n}{dx^n}(e^{-x}x^n)$$

的零点,而系数 A_i 为

$$A_i=\frac{(n!)^2}{x_i[L_n'(x_i)]^2}.$$

在表中的数 x_i 及 A_i 可用.

Gauss-Laguerre 公式的推导要利用 Laguerre 多项式的性质,与 Gauss-Legendre 公式十分相似.

3. **Gauss-Hermite 公式**取形式

$$\int_{-\infty}^{\infty} e^{-x^2} y(x) dx \approx \sum_{i=1}^{n} A_i y(x_i),$$

自变量 x_i 为 n 次 Hermite 多项式

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}(e^{-x^2})$$

的零点,而系数 A_i 为

$$A_i = \frac{2^{n+1} n! \sqrt{\pi}}{[H_n'(x_i)]^2}.$$

在表中的数 x_i 及 A_i 可用.

4. **Gauss-Chebyshev 公式**取形式

$$\int_{-1}^{1} \frac{y(x)}{\sqrt{1-x^2}} dx \approx \frac{\pi}{n} \sum_{i=1}^{n} y(x_i),$$

自变量 x_i 为 n 次 Chebyshev 多项式 $T_n(x) = \cos(n \arccos x)$的零点.

题　解

Gauss 方法

15.1 对一个逼近 $y(x)$的密切多项式在自变量 x_1 到 x_n 处积分 Hermite 公式.

解 ☞ 这里为方便起见,在我们的密切多项式中删去自变量 x_0.这一要求只是对我们在第十章中的公式稍加改变.Hermite 公式本身变成

$$p(x) = \sum_{i=1}^{n} [1 - 2L_i'(x_i)(x - x_i)][L_i(x)]^2 y_i + (x - x_i)[L_i(x)]^2 y_i',$$

其中 $L_i(x) = F_i(x)/F_i(x_i)$为 Lagrange 乘算子函数, $F_i(x)$是乘积 $F_i(x) = \prod_{k \neq i} (x - x_k)$.

积分后得到

$$\int_a^b w(x) p(x) dx = \sum_{i=1}^{n} (A_i y_i + B_i y_i'),$$

其中

$$A_i = \int_a^b w(x)[1 - 2L_i'(x_i)(x - x_i)][L_i(x)]^2 dx,$$

$$B_i = \int_a^b w(x)(x - x_i)][L_i(x)]^2 dx.$$

15.2 找出题 15.1 中公式的截断误差.

解 ☞ 十足地令人惊讶的是,这样来得到公式比从简单配置多项式来得到公式更为容易,因为可以直接地应用中值定理.因为我们删去了一个自变量,在 $n+1$ 的地方换成 n 后 Hermite 公式的误差(题 10.4)变成

$$y(x) - p(x) = \frac{y^{(2n)}(\xi)}{(2n)!}[\pi(x)]^2.$$

乘以 $w(x)$并加以积分得

$$\int_a^b w(x)[y(x) - p(x)] dx = \int_a^b w(x) \frac{y^{(2n)}(\xi)}{(2n)!}[\pi(x)]^2 dx.$$

由于 $w(x)$选成为一个非负函数而$[\pi(x)]^2$ 当然是正的,所以由中值定理立得截断误差为

$$E = \int_a^b w(x)[y(x) - p(x)] dx = \frac{y^{(2n)}(\theta)}{(2n)!} \int_a^b w(x)[\pi(x)]^2 dx.$$

这里 $a < \theta < b$,而 θ 如通常那样是无从可知的.注意若 $y(x)$为一个次数不超过 $2n-1$ 的多项式,这个误差就会精确地为零.我们的公式对所有这类多项式为精确的.

15.3 证明所有系数 B_i 将为零,若

$$\int_a^b w(x)\pi(x)x^k dx = 0, \qquad 当\ k = 0, 1, \cdots, n-1.$$

证 ☞ 由题 8.3 $(x-x_i)L_i(x)=\pi(x)/\pi'(x_i)$. 把它代入 B_i 的公式

$$B_i=\frac{1}{\pi'(x_i)}\int_a^b w(x)\pi(x)L_i(x)\mathrm{d}x.$$

但是, $L_i(x)$ 是 x 的 $n-1$ 次多项式, 故

$$B_i=\frac{1}{\pi'(x_i)}\int_a^b w(x)\pi(x)\sum_{k=0}^{n-1}\alpha_k x^k\mathrm{d}x$$

$$=\frac{1}{\pi'(x_i)}\sum_{k=0}^{n-1}\alpha_k\int_a^b w(x)\pi(x)x^k\mathrm{d}x=0.$$

15.4 定义正交函数并且以正交性重新叙述题 15.3 的结果.

解 ☞ 函数 $f_1(x)$ 及 $f_2(x)$ 称为在区间 (a,b) 上对权函数 $w(x)$ 来说是正交的, 若

$$\int_a^b w(x)f_1(x)f_2(x)\mathrm{d}x=0.$$

我们公式中的系数 B_i 将会是零, 若 $\pi(x)$ 正交于 x^p, 当 $p=0,1,\cdots,n-1$. 进一步, $\pi(x)$ 将正交于不超过 $n-1$ 次的任何多项式, 包括 Lagrange 乘算子函数 $L_i(x)$. 这种正交性依赖于并且决定我们对配置点 x_i 的选择, 在本章的余下部分都作此假设.

15.5 证明随所有的 $B_i=0$, 系数 A_i 还原为

$$A_i=\int_a^b w(x)[L_i(x)]^2\mathrm{d}x,$$

且因此为正数.

证 ☞

$$A_i=\int_a^b w(x)[L_i(x)]^2\mathrm{d}x-2L_i'(x_i)B_i,$$

当 $B_i=0$ 时还原为所求的形式.

15.6 导出较简单的公式 $A_i=\int_a^b w(x)L_i(x)\mathrm{d}x$.

解 ☞ 假如我们可以证明 $\int_a^b w(x)L_i(x)[L_i(x)-1]\mathrm{d}x=0$ 就得这个结果.

但是, $L_i(x)-1$ 必定含有 $(x-x_i)$ 作为一个因子, 因为 $L_i(x_i)-1=1-1=0$. 因此

$$L_i(x)[L_i(x)-1]=\frac{\pi(x)}{\pi'(x_i)(x-x_i)}[L_i(x)-1]$$

$$=\pi(x)p(x)$$

具有最多为 $n-1$ 次的 $p(x)$. 于是题 15.3 保证该积分为零.

15.7 本节的积分公式现在可以写成

$$\int_a^b w(x)y(x)\mathrm{d}x\approx\sum_{i=1}^n A_i y(x_i).$$

其中 $A_i=\int_a^b w(x)L_i(x)\mathrm{d}x$, 而自变量值 x_i 是由题 15.3 中所要求的正交性来选定的. 这个公式是由积分一个在自变量 x_i 处取值 y_i 及 y_i' 的 $2n-1$ 次密切多项式而得到的. 证明同样的公式可以通过对更为简单的单由 y_i 值决定的 $n-1$ 次配置多项式作积分而得到. (这是看待 Gauss 公式的一种方法, 它们从相对低次的多项式来达到高的精确度.)

证 ☞ 配置多项式为 $p(x)=\sum_{i=1}^n L_i(x)y(x_i)$, 故积分产生

$$\int_a^b w(x)p(x)\mathrm{d}x=\sum_{i=1}^n A_i y(x_i),$$

正如所提示的那样. 这里 $p(x)$ 代表配置多项式. 在题 15.1 中它是更为复杂的密切多项式. 二者均导出同样的积分公式. (关于它的一个特定例子, 见题 15.25.)

Gauss-Legendre 公式

15.8 特殊情况 $w(x)=1$ 导出 Gauss-Legendre 公式. 习惯上使用积分区间 $(-1,1)$. 作为一

个初步的练习，直接由题 15.3 的条件对值 $n=3$ 来确定自变量 x_k，

$$\int_{-1}^{1}\pi(x)x^k\mathrm{d}x=0\qquad k=0,1,\cdots,n-1.$$

解 ☞ 此时多项式 $\pi(x)$ 为三次的，不妨设 $\pi(x)=a+bx+cx^2+x^3$，积分可得

$$2a+\frac{2}{3}c=0,\quad \frac{2}{3}b+\frac{2}{5}=0,\quad \frac{2}{3}a+\frac{2}{5}c=0.$$

由它们很快地推出 $a=c=0,b=-\frac{3}{5}$. 这使得

$$\pi(x)=x^3-\frac{3}{5}x=\left(x+\sqrt{\frac{3}{5}}\right)x\left(x-\sqrt{\frac{3}{5}}\right).$$

因此配置自变量为 $x_k=-\sqrt{\frac{3}{5}},0,\sqrt{\frac{3}{5}}$.

理论上讲，此法对任何 n 值均可产生 x_k，然而可用一个更为精巧的方法使其更为快捷.

15.9　n 次 Legendre 多项式定义为

$$P_n(x)=\frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n.$$

取 $P_0(x)=1$，证明对 $k=0,1,\cdots,n-1$ 有

$$\int_{-1}^{1}x^kP_n(x)\mathrm{d}x=0,$$

从而 $P_n(x)$ 也正交于任何小于 n 次的多项式.

解 ☞ 应用分部积分 k 次得

$$\int_{-1}^{1}x^k\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n\mathrm{d}x=\underbrace{\left[x^k\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}(x^2-1)^n\right]_{-1}^{1}}_{=0}$$

$$-\int_{-1}^{1}kx^{k-1}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}(x^2-1)^n\mathrm{d}x$$

$$=\cdots=(-1)^k k!\int_{-1}^{1}\frac{\mathrm{d}^{n-k}}{\mathrm{d}x^{n-k}}(x^2-1)^n\mathrm{d}x=0.$$

15.10　证明 $\int_{-1}^{1}x^nP_n(x)\mathrm{d}x=\frac{2^{n+1}(n!)^2}{(2n+1)!}$.

证 ☞ 在上题中取 $k=n$，

$$\int_{-1}^{1}x^n\frac{\mathrm{d}x}{\mathrm{d}x^n}(x^2-1)^n\mathrm{d}x=(-1)^n n!\int_{-1}^{1}(x^2-1)^n\mathrm{d}x$$

$$=2n!\int_0^1(1-x^2)^n\mathrm{d}x=2n!\int_0^{\pi/2}\cos^{2n+1}t\,\mathrm{d}t.$$

最后的积分可处理如下：

$$\int_0^{\pi/2}\cos^{2n+1}t\,\mathrm{d}t=\underbrace{\left[\frac{\cos^{2n}t\sin t}{2n+1}\right]_0^{\pi/2}}_{=0}+\frac{2n}{2n+1}\int_0^{\pi/2}\cos^{2n-1}t\,\mathrm{d}t$$

$$=\cdots=\frac{2n(2n-1)\cdots2}{(2n+1)(2n-1)\cdots3}\int_0^{\pi/2}\cos t\,\mathrm{d}t.$$

所以

$$\int_{-1}^{1}x^n\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n\mathrm{d}x=2n!\frac{2n(2n-2)\cdots2}{(2n+1)(2n-1)\cdots3}.$$

现在以 $2n(2n-2)\cdots2=2^nn!$ 乘分子与分母并记住 $P_n(x)$ 的定义就得到所求

$$\int_{-1}^{1}x^nP_n(x)\mathrm{d}x=\frac{1}{2^nn!}2n!\frac{2^nn!2^nn!}{(2n+1)!}$$

$$=\frac{2^{n+1}(n!)^2}{(2n+1)!}.$$

15.11　证明 $\int_{-1}^{1}[P_n(x)]^2\mathrm{d}x=\frac{2}{2n+1}$.

证 ☞ 在一个 $P_n(x)$ 因子中将 x 的最高次幂分离出来

$$\int_{-1}^{1}[P_n(x)]^2\mathrm{d}x = \int_{-1}^{1}\left[\frac{1}{2^n n!}\frac{2n!}{n!}x^n+\cdots\right]P_n(x)\mathrm{d}x.$$

由题 15.9，x^n 以下的幂对积分不作贡献，再用上题结果，我们有

$$\int_{-1}^{1}[P_n(x)]^2\mathrm{d}x = \frac{(2n)!}{2^n(n!)^2}\cdot\frac{2^{n+1}(n!)^2}{(2n+1)!} = \frac{2}{2n+1}.$$

15.12 证明当 $m\neq n$ 时，$\int_{-1}^{1}P_m(x)P_n(x)\mathrm{d}x = 0$.

证 ☞ 写出较低次多项式，我们发现它当中的每次幂项都正交于较高次的多项式．特别取 $m=0$ 及 $n\neq 0$ 时我们有特殊情况 $\int_{-1}^{1}P_n(x)\mathrm{d}x = 0$.

15.13 证明 $P_n(x)$ 在 -1 与 1 之间有 n 个实零点.

证 ☞ 多项式 $(x^2-1)^n$ 为 $2n$ 次，它在 ±1 处有多重零点．因此由 Rolle 定理它的导数在区间内部有零点．一阶导数在 ±1 处也为零．因此总共有三个零点．二阶导数由 Rolle 定理保证有二个内部零点．它于 ±1 处也为零，因此总共四个零点．以这种方式进行下去我们发现 n 阶导数由 Rolle 定理保证有 n 个内部零点．除了一个常数因子外这个导数就是 Legendre 多项式 $P_n(x)$.

15.14 证明对于权函数 $w(x)=1$，$\pi(x)=[2^n(n!)^2/(2n)!]P_n(x)$.

证 ☞ 将 $P_n(x)$ 的零点称作 $x_1,\cdots x_n$．于是

$$\left[\frac{2^n n!}{(2n)!}\right]P_n(x) = (x-x_1)\cdots(x-x_n).$$

对 $\pi(x)$ 的惟一的其他要求是它与 x^k 正交 $k=0,1,\cdots,n-1$．而这一点从题 15.9 得出.

15.15 直接从定义计算头几个 Legendre 多项式，注意在任一个此类多项式中仅有偶数幂或奇数幂出现.

解 ☞ $P_0(x)$ 定义为 1，于是我们得到

$$P_1(x) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}(x^2-1) = x,$$

$$P_2(x) = \frac{1}{8}\frac{\mathrm{d}^2}{\mathrm{d}x^2}(x^2-1)^2 = \frac{1}{2}(3x^2-1),$$

$$P_3(x) = \frac{1}{48}\frac{\mathrm{d}^3}{\mathrm{d}x^3}(x^2-1)^3 = \frac{1}{2}(5x^2-3x),$$

$$P_4(x) = \frac{1}{16\cdot 24}\frac{\mathrm{d}^4}{\mathrm{d}x^4}(x^2-1)^4 = \frac{1}{8}(35x^4-30x^2+3).$$

类似地有

$$P_5(x) = \frac{1}{8}(63x^5-70x^3+15x),$$

$$P_6(x) = \frac{1}{16}(231x^6-315x^4+105x^2-5),$$

$$P_7(x) = \frac{1}{16}(429x^7-693x^5+315x^3-35x),$$

$$P_8(x) = \frac{1}{128}(6435x^8-12{,}012x^6+6930x^4-1260x^2+35)$$

等等．由于 $(x^2-1)^n$ 只含 x 的偶次幂，微分 n 次的结果将只含有偶次项或奇次项.

15.16 证明 x^n 可以被表示成 Legendre 多项式直到 $P_n(x)$ 的组合．同样对任意 n 次的多项式也成立.

证 ☞ 对逐次幂依次求解，我们得到

$$1 = P_0(x),\qquad x = P_1(x),\qquad x^2 = \frac{1}{3}[2P_2(x)+P_0(x)],$$

$$x^3 = \frac{1}{5}[2P_3(x)+3P_1(x)],\qquad x^4 = \frac{1}{35}[8P_4(x)+20P_2(x)+7P_0(x)]$$

等等. 每个 $P_k(x)$ 都从一个 x^k 的非零项开始的, 这个事实允许这个过程可以不断地继续下去.

15.17　证明 Legendre 多项式的递推公式

$$(n+1)P_{n+1}(x)=(2n+1)xP_n(x)-nP_{n-1}(x).$$

证　☞ 多项式 $xP_n(x)$ 为 $n+1$ 次的, 于是(参看题 15.16)可以被表成组合式

$$xP_n(x)=\sum_{i=0}^{n+1}c_iP_i(x),$$

乘以 $P_k(x)$ 并积分得到

$$\int_{-1}^{1}xP_k(x)P_n(x)\mathrm{d}x=c_k\int_{-1}^{1}P_k^2(x)\mathrm{d}x.$$

在右侧的所有的其他项由于下同次数的 Legendre 多项式是正交的. 但是当 $k<n-1$ 时我们知道 $P_n(x)$ 也与 $xP_k(x)$ 正交, 因为此时这个乘积最多为 $n-1$ 次的(参见题 15.9). 这使得当 $k<n-1$ 时 $c_k=0$ 且

$$xP_n(x)=c_{n+1}P_{n+1}(x)+c_nP_n(x)+c_{n-1}P_{n-1}(x).$$

注意到, 由定义 $P_n(x)$ 中 x^n 的系数是 $(2n)!/2^n(n!)^2$, 我们比较上式中 x^{n+1} 的系数得到

$$\frac{(2n)!}{2^n(n!)^2}=c_{n+1}\frac{(2n+2)!}{2^{n+1}[(n+1)!]^2}.$$

由此得 $c_{n+1}=(n+1)/(2n+1)$. 比较 x^n 的系数并记住在任何 Legendre 多项式中只有非偶即奇的幂出现, 故 $c_n=0$. 返回到我们的积分来决定 c_{n-1}, 取 $k=n-1$ 时我们想象 $P_n(x)$ 以幂的和写出. 只有 x^{n-1} 才需要予以考虑, 因为较低的项, 即使乘以 x 还是正交于 $P_n(x)$. 这就导出了

$$\frac{(2n-2)!}{2^{n-1}[(n-1)!]^2}\int_{-1}^{1}x^nP_n(x)\mathrm{d}x=c_{n-1}\int_{-1}^{1}P_{n-1}^2(x)\mathrm{d}x,$$

并用题 15.10 及 15.11 的结果人们容易发现 $c_{n-1}=n/(2n+1)$. 将这些系数代入我们关于 $xP_n(x)$ 的表达式便获得所要的递推公式. 作为额外的收获我们还有

$$\int_{-1}^{1}xP_{n-1}(x)P_n(x)\mathrm{d}x=\frac{n}{2n+1}\frac{2}{2n-1}=\frac{2n}{4n^2-1}.$$

15.18　举例说明递推公式的使用.

解　☞ 取 $m=5$, 我们得到

$$P_6(x)=\frac{11}{6}xP_5(x)-\frac{5}{6}P_4(x)$$
$$=\frac{1}{16}(231x^6-315x^4+105x^2-5).$$

而取 $m=6$ 时,

$$P_7(x)=\frac{13}{7}xP_6(x)-\frac{6}{7}P_5(x)$$
$$=\frac{1}{16}(429x^7-693x^5+315x^3-35x),$$

确认了题 15.15 中所得到的结果. 这递推过程很适合于这些多项式的自动计算, 而题 15.15 中的微分过程并不如此.

15.19　导出 Christoffel 恒等式,

$$(t-x)\sum_{i=1}^{\infty}(2i+1)P_i(x)P_i(t)=(n+1)[P_{n+1}(t)P_n(x)-P_n(t)P_{n+1}(x)].$$

解　☞ 题 15.17 中的递推公式可以通过乘以 $P_i(t)$ 得到

$$(2i+1)xP_i(x)P_i(t)=(i+1)P_{i+1}(x)P_i(t)-iP_{i-1}(x)P_i(t).$$

再将自变量 x 与 t 互换再写一遍(因为它对任意的 x 与 t 都是成立的)并令之相减我们得到

$$(2i+1)(t-x)P_i(x)P_i(t)=(i+1)[P_{i+1}(t)P_i(x)-P_i(t)P_{i+1}(x)]$$
$$-i[P_i(t)P_{i-1}(x)-P_{i-1}(t)P_i(x)].$$

将 i 从 1 到 n 求和, 并注意到在右边的前后相抵销, 我们有

$$(t-x)\sum_{i=1}^{n}(2i+1)P_i(x)P_i(t)=(n+1)$$

$$\cdot [P_{n+1}(t)P_n(x) - P_n(t)P_{n+1}(x)] - (t - x),$$

最后一项可以移到左边,还可以被吸收到求和记号中作为 $i=0$ 这一项.这就是 Christoffel 恒等式.

15.20 用 Christoffel 恒等式计算关于 Gauss-Legendre 公式的积分系数,证明

$$A_k = \frac{2}{nP_n'(x_k)P_{n-1}(x_k)}.$$

证 ☞ 令 x_k 为 $P_n(x)$的一个零点.然后在上题中用 x_k 置换 t,就成了

$$\frac{(n+1)P_{n+1}(x_k)P_n(t)}{x - x_k} = -\sum_{i=0}^{n}(2i+1)P_i(x)P_i(x_k).$$

现在从 -1 积分到 1.由题 15.12 中的特殊情况,只有 $i=0$ 项还留在右侧,从而我们有

$$\int_{-1}^{1}\frac{P_n(x)}{x - x_n}dx = \frac{-2}{(n+1)P_{n+1}(x_k)}.$$

取 $x = x_k$ 时递推公式变成$(n+1)P_{n+1}(x_k) = -nP_{n-1}(x_k)$,它允许我们将上式改写成

$$\int_{-1}^{1}\frac{P_n(x)}{x - x_k}dx = \frac{2}{nP_{n-1}(x_k)}.$$

由题 15.6 及 15.14 我们现在得到

$$A_k = \int_{-1}^{1}L_k(x)dx = \int_{-1}^{1}\frac{\pi(x)}{\pi'(x_k)(x - x_k)}dx$$
$$= \int_{-1}^{1}\frac{P_n(x)}{P_n'(x_k)(x - x_k)}dx,$$

立刻导出所述的结果.

15.21 证明$(1 - x^2)P_n'(x) + nxP_n(x) = nP_{n-1}(x)$,它对简化题 15.20 的结果有用.

证 ☞ 我们首先注意组合$(1 - x^2)P_n' + nxP_n$ 最多为 $n+1$ 次的.然而,以 A 表示$P_n(x)$的最高次项系数,容易发现 x^{n+1}为$(-nA + nA)$所乘.所以它不包含在内.由于 P_n 不含有 x^{n-1}项,故在我们的组合中也就不含 x^n 项.它的次数最多为$n-1$,且由题 15.16 它可以表示为

$$(1 - x^2)P_n'(x) + nxP_n(x) = \sum_{i=0}^{n-1}c_iP_i(x).$$

像我们在递推公式的推导中那样进行,我们现在用 $P_k(x)$去乘并且积分.由于正交性在右侧只有第 k 项不为零的,我们得到

$$\frac{2}{2k+1}c_k = \int_{-1}^{1}(1 - x^2)P_n'(x)P_k(x)dx$$
$$+ n\int_{-1}^{1}xP_n(x)P_k(x)dx.$$

对第一项进行分部积分,积出来的那一部分则由于有因子$(1 - x^2)$因而为零.由此得到

$$\frac{2}{2k+1}c_k = -\int_{-1}^{1}P_n(x)\frac{d}{dx}[(1 - x^2)P_k(x)]dx$$
$$+ n\int_{-1}^{1}xP_n(x)P_k(x)dx.$$

当 $k < n-1$ 时二个被积函数都是 $P_n(x)$乘上一个次数不大于 $n-1$ 的多项式.由题 15.9 所有这类 c_k 将是零.当 $k = n-1$ 时,最后一积分由题 15.17 的附加结果得到.在第一个积分中只有 $P_{n-1}(x)$的最高项有贡献(还是由于题 15.9)使这一项为

$$\int_{-1}^{1}P_n(x)\frac{d}{dx}\left\{x^2\frac{(2n-2)!}{2^{n-1}[(n-1)!]^2}x^{n-1}\right\}dx.$$

利用题 15.10,它现在简化为

$$\frac{(2n-2)!}{2^{n-1}[(n-1)!]^2}\cdot(n+1)\frac{2^{n+1}(n!)^2}{(2n+1)!} = \frac{2n(n+1)}{(2n+1)(2n-1)}.$$

将这些不同结果代入,我们得到

$$c_{n-1} = \frac{2n-1}{2}\left[\frac{2n(n+1)}{(2n+1)(2n-1)} + \frac{2n^2}{(2n+1)(2n-1)}\right] = n.$$

证明完毕.

15.22 应用题 15.21 来获得 $A_k = \dfrac{2(1 - x_k^2)}{n^2[P_{n-1}(x_k)]^2}$.

解 ☞ 令 $x = x_k$ 为 $P_n(x)$ 的一个零点，我们得到 $(1-x_k^2)P_n'(x_k) = nP_{n-1}(x_k)$. 含导数的那个因子现可以换成我们在题 15.20 中的结果，就产生所求的结果.

15.23 Gauss-Legendre 积分公式现在可以表示为

$$\int_{-1}^{1} y(x)\,\mathrm{d}x \approx \sum_{i=1}^{n} A_i y(x_i),$$

此处 x_k 为 $P_n(x)$ 的零点，而系数 A_k 为题 15.22 所给出的. 将这些数对 $n = 2, 4, 6, \cdots, 16$ 列表.

解 ☞ 当 $n=2$ 时我们解 $P_2(x) = \frac{1}{2}(3x^2-1) = 0$ 便得到 $x_k = \pm\sqrt{\frac{1}{3}} = \pm 0.57735027$. 二个系数证明为一样的. 题 15.22 使 $A_k = 2\left(1-\frac{1}{3}\right)\Big/\left[4\cdot\frac{1}{3}\right] = 1$

当 $n=4$ 时我们解 $P_4(x) = \frac{1}{8}(35x^4-30x^2+3) = 0$ 便得到 $x_k^2 = (15 \pm 2\sqrt{30})/35$. 导出四个点为 $x_k = \pm\left[(15 \mp 2\sqrt{30})\big/35\right]^{1/2}$.

计算它们并将其插入题 15.22 的公式中便得在表 15.1 中给出的 x_k, A_k. 对于更大一些的 n 这些结果可以同样的方式获得，高次多项式的零点是用熟悉的 Newton 逐次逼近法获得的（这个方法出现在较后的章中）.

表 15.1

n	x_k	A_k	n	x_k	A_k
2	±0.57735027	1.00000000	12	±0.58731795	0.20316743
4	±0.86113631	0.34785485		±0.36783150	0.23349254
	±0.33998104	0.65214515		±0.12533341	0.24914705
6	±0.93246951	0.17132449	14	±0.98628381	0.03511946
	±0.66120939	0.36076157		±0.92843488	0.08015809
	±0.23861919	0.46791393		±0.82720132	0.12151857
8	±0.96028986	0.10122854		±0.68729290	0.15720317
	±0.79666648	0.22238103		±0.51524864	0.18553840
	±0.52553241	0.31370665		±0.31911237	0.20519846
	±0.18343464	0.36268378		±0.10805495	0.21526385
10	±0.97390653	0.06667134	16	±0.98940093	0.02715246
	±0.86506337	0.14945135		±0.94457502	0.06225352
	±0.67940957	0.21908636		±0.86563120	0.09515851
	±0.43339539	0.26926672		±0.75540441	0.12462897
	±0.14887434	0.29552422		±0.61787624	0.14959599
12	±0.98156063	0.04717534		±0.45801678	0.16915652
	±0.90411725	0.10693933		±0.28160355	0.18260342
	±0.76990267	0.16007833		±0.09501251	0.18945061

15.24 应用两点公式于 $\int_0^{\pi/2} \sin t\,\mathrm{d}t$.

解 ☞ 变量变换 $t = \pi(x+1)/4$ 将它转化为我们的标准区间如下

$$\int_{-1}^{1} \frac{\pi}{4}\sin\frac{\pi(x+1)}{4}\,\mathrm{d}x.$$

而 Gauss 节点 $x_k = \pm 0.57735027$ 导出 $y(x_1) = 0.32589$, $y(x_2) = 0.94541$. 现在两点公式保证 $(\pi/4)(0.32589 + 0.94541) = 0.99848$, 它几乎准确到 3 位. 两点 Gauss 公式产生比用 7 点的梯形公式更好的结果. 误差为千分之 2!

来考虑一下一点公式能得到什么是颇有兴趣的事. 当 $n=1$ 时 Gauss-Legendre 结果是（正如人

们容易验证的那样) $\int_{-1}^{1} y(x)dx \approx 2y(0)$. 对于正弦函数它就是

$$\int_{-1}^{1}\frac{\pi}{4}\sin\frac{\pi(x+1)}{4}dx \approx \frac{\pi}{4}\sqrt{2} \approx 1.11$$

它准确到大约百分之十以内.

15.25 通过展示公式所基于的多项式,来解释在题 15.24 中所用的特别简单公式的精确度.

解 ☞ $n=1$ 的公式可以通过积分零次配置多项式 $p(x)=y(x_1)=y(0)$ 来得到. 然而,它也可以从 $2n-1=1$ 次密切多项式得到,这就是 Gauss 方法的思想. 由 Hermite 公式多项式为 $y(0)+xy'(0)$. 在 -1 与 1 之间积分这个线性函数其结果同样为 $2y(0)$,导数项不作贡献. 零次配置多项式产生与一次多项式相同的结果,因为配置点就是 Gauss 点(图 15.1).

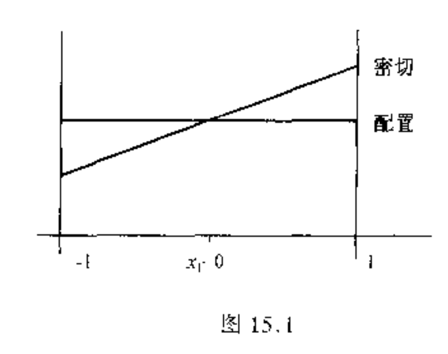


图 15.1

图 15.2

类似地,$n=2$ 公式可以通过积分一次多项式而得到,配置点为 Gauss 点

$$\int_{-1}^{1}\left(\frac{x-r}{-2r}y_1+\frac{x+r}{2r}y_2\right)dx=y_1+y_2,$$

其中 $r=\sqrt{\frac{1}{3}}$. 通过积分 3 次密切多项式可以得到同样的结果,因为

$$\int_{-1}^{1}\left[\left(1+\frac{x+r}{r}\right)\frac{3}{4}(x-r)^2y_1+\left(1-\frac{x-r}{r}\right)\frac{3}{4}(x+r)^2y_2 + \frac{3}{4}(x^2-r^2)(x-r)y_1'+\frac{3}{4}(x^2-r^2)(x+r)y_2'\right]dx=y_1+y_2.$$

一次多项式完成得这样好,源于配置点为 Gauss 点(图 15.2).

15.26 应用 Gauss 4 点公式于题 15.24 的积分.

解 ☞ 用同样的自变量变换,4 点公式产生 $\sum_{i=1}^{4}A_iy_i=1.000000$,准确到 6 位. 与 Simpson32-点公式的结果 1.0000003 以及 Simpson64-点的结果 0.99999983 相比,我们发现它优于这两者.

15.27 将题 15.2 的截断误差估计用于 Gauss-Legendre 逼近公式的特殊情况.

解 ☞ 综合题 15.2、15.11 及 15.14,我们得到误差为

$$E=\frac{y^{(2n)}(\theta)}{(2n)!}\left[\frac{2^n(n!)^2}{(2n)!}\right]^2\frac{2}{2n+1}=\frac{2^{2n+1}(n!)^4}{(2n+1)[(2n)!]^3}y^{(2n)}(\theta).$$

假如 $y(x)$ 的导数难以计算的话,这就不是一个便于应用的公式. 然而,通过对小的 n 计算 $y^{(2n)}$ 的系数,对 Gauss 公式之精度还可作些进一步的探讨.

$$n=2 \qquad E=0.0074y^{(4)},$$
$$n=4 \qquad E=0.0000003y^{(8)},$$
$$n=6 \qquad E=1.5(10^{-12})y^{(12)}.$$

15.28 应用题 15.27 的误差估计于题 15.24 的积分,并将其与实际误差进行比较.

解 ☞ 在进行一个将积分变成标准形式的变量变换之后,我们得到

$$|y^{(4)}(x)|<\left(\frac{\pi}{4}\right)^5, \qquad |y^{(8)}(x)|<\left(\frac{\pi}{4}\right)^9.$$

对 $n=2$ 它使我们的误差估计为 $E=(0.0074)(0.298)=0.00220$,同时对 $n=4$ 我们得到 $E=(0.0000003)(0.113)=0.00000003$. 实际误差为 0.00152,并且准确到 6 位时为零. 因此我们的估计与我们的结果相容.

此例提供一种有利的情况.即使用近似方法,正弦函数仍易于积分.因为它的导数都是囿于相同的常数,即 $\pi/4$ 的幂肯定随着变量变换而进入,但是在这种情况下它们实际上是带来益处.下个例子处理一个常用函数,它的导数就没有如此有利的性质.

15.29　应用 Gauss-Legendre 公式于 $\int_0^{\pi/4}\log(1+t)\mathrm{d}t$.

解　☞ 这个积分的准确值为

$$\left(1+\frac{\pi}{2}\right)\left[\log\left(1+\frac{\pi}{2}\right)-1\right]-1=0.856590,$$

精确到 6 位.变量变换 $t=\pi(x+1)/4$ 将积分转换成

$$\int_{-1}^{1}\frac{\pi}{4}\log\left[1+\frac{\pi(x+1)}{4}\right]\mathrm{d}x.$$

这个新被积函数的 4 阶导数为 $(\pi/4)^5[-6/(1+t)^4]$.在积分区间内它不会超过 $6(\pi/4)^5$,所以截断误差不会超过 $6(\pi/4)^5(0.0074)$.假如我们用 2 点 Gauss 公式的话,这是正弦函数积分的相应估计的 6 倍.类似地,8 阶导数为 $(\pi/4)^9[-7!/(1+t)^8]$.这意味着一个截断误差最多 $(\pi/4)^9\cdot 7!(0.0000003)$,它 7! 倍于正弦函数积分的相应估计.当逐阶的正弦函数的导数都保持以 1 为界时,那些对数函数的导数却以阶乘式增长.差分对任何一种公式截断误差都有明显的影响.或许特别是对那些含有特别高阶导数的 Gauss 公式误差的影响.即使如此,这些公式还是表现良好.使用 2 点公式我们得到 0.858,而 4 点公式的结果为0.856592,它在最后一位只差二个单位.6 点 Gauss 公式效果甚佳,精确到 6 位,虽然它的截断误差项包含了 $y^{(12)}(x)$.(其大小近似也为 12!).相反,Simpson 法则需要 64 点才能得出相同的 6 位结果.

函数 $\log(1+t)$ 在 $t=-1$ 处有一个奇异点.它并不在积分区间上,但是接近,而即使一个复奇异点在近处也能造成明显的慢类型的收敛.

15.30　积分区间的长度是怎样影响 Gauss 公式的?

解　☞ 对一个在区间 $a\leqslant t\leqslant b$ 上的积分,变量变换 $t=a+\frac{b-a}{2}(x+1)$ 将产生标准区间 $-1\leqslant x\leqslant 1$.它还使得

$$\int_a^b y(t)\mathrm{d}t=\int_{-1}^{1}\frac{b-a}{2}y\left[a+\frac{b-a}{2}(x+1)\right]\mathrm{d}x$$

对截断误差的影响是在导数因子上.它是

$$\left(\frac{b-a}{2}\right)^{2n+1}y^{(2n)}(t).$$

在那些给定的 $b-a$ 正好为 $\pi/2$ 的例子中,这个区间长度实际上有助于减少误差,然而对一个较长的区间来说 $b-a$ 幂的存在显然要放大误差.

15.31　应用 Gauss 方法于 $(2/\sqrt{\pi})\int_0^4 e^{-t^2}\mathrm{d}t$.

解　☞ 这个误差函数的较高阶导数不容易现实地进行估计.往前计算,人们发现 $n=4,6,8,10$ 时的公式给出的结果为:

n	4	6	8	10
近似值	0.986	1.000258	1.000004	1.000000

对于更大的 n 其结果与 $n=10$ 的相同,它表明精确到 6 位.我们已经在 Taylor 级数的一个有意义的应用中计算过这个积分(题 14.32),而且发现它等于 1,准确到 6 位.作为比较,为了达到 6 位精度 Simpson 公式需要 32 点.

15.32　应用 Gauss 方法于 $\int_0^4\sqrt{1+\sqrt{t}}\,\mathrm{d}t$.

解　☞ $n=4,8,12,16$ 时的公式给出的结果为

n	4	8	12	16
近似值	6.08045	6.07657	6.07610	6.07600

它表明精确到4位.可以通过变量变换得到精确积分为$\frac{8}{5}\left(2\sqrt{3}+\frac{1}{3}\right)$.它等于6.07590准确到5位.我们观察到这里所得到的精度与上题相比是相形见拙的.其原因是我们的平方根被积函数不像指数函数那样光滑.它的高阶导数增长很大,像阶乘那样.我们其他的公式也能感受到这些大导数的影响.例如用 Simpson 法则产生这些值:

点数	16	64	256	1024
Simpson 值	6.062	6.07411	6.07567	6.07586

即使用一千点它也不可能设法达到如上题中只用32点就能达到的精度.

15.33 导出对 Gauss 公式截断误差的 Lanczos 估计.

解 ☞ 关系式$\int_{-1}^{1}[xy(x)]'\mathrm{d}x=y(1)+y(-1)$精确地成立.令$I$为由$n$-点 Gauss 公式所得的$y(x)$的近似积分,而$I^*$为关于$[xy(x)]'$的相应结果.由于$[xy(x)]'=y(x)+xy'(x)$,

$$I^*=I+\sum_{i=1}^{n}A_ix_iy'(x_i).$$

所以I^*中的误差为

$$E^*=y(1)+y(-1)-I-\sum_{i=1}^{n}A_ix_iy'(x_i).$$

称I本身的误差为E,我们知道

$$E=C_ny^{(2n)}(\theta_1),\qquad E^*=C_n(xy)^{(2n+1)}(\theta_2),$$

对介于-1与1之间的适当的θ_1及θ_2.假设$\theta_1=\theta_2=0$.一方面$(xy)^{(2n+1)}(0)/(2n)!$是$(xy)'$的 Taylor 级数展式中x^{2n}的系数,而另一方面

$$y(x)=\cdots+\frac{y^{(2n)}(0)x^{2n}}{(2n)!}+\cdots$$

直接导出
$$[xy(x)]'=\cdots+\frac{(2n+1)y^{(2n)}(0)x^{2n}}{(2n)!}+\cdots.$$

由它我们推出
$$(xy)^{(2n+1)}(0)=(2n+1)y^{(2n)}(0).$$

因此近似地有$E^*=(2n+1)E$,使得

$$E\approx\frac{1}{2n+1}\left[y(1)+y(-1)-I-\sum_{i=1}^{n}A_ix_iy'(x_i)\right].$$

它包含了将 Gauss 公式应用于$xy'(x)$以及$y(x)$本身,但是它避免了常常是麻烦的$y^{(2n)}(x)$计算.令$\theta_1=\theta_2=0$是在推导这个公式中关键的一步.它被发现对于诸如题15.31中的光滑被积函数比具有高阶导数的被积函数更为合理,而它看来是合理的,则由于当$y^{(2n+1)}$为小量时$y^{(2n)}(\theta_1)/y^{(2n)}(\theta_2)$必是接近于1.

15.34 应用上题的误差估计于题15.31的积分.

解 ☞ 当$n=8$时 Lanczos 估计为0.000004恒等于真正的误差.当$n=10$以上的情况,Lanczos 估计准确地预测一个到6位为零的误差.然而,假如它应用于题15.32的积分,那里被积函数十分不光滑.Lanczos 估计证明是用起来太保守.该公式的适用性范围仍是一个待定问题.

其他 Gauss 型积分

15.35 什么是 Gauss-Laguerre 公式?

解 ☞ 关于近似积分的这些公式具有形式

$$\int_0^{\infty}e^{-x}y(x)\mathrm{d}x\approx\sum_{i=1}^{n}A_iy(x_i),$$

自变量值 x_i 为 n 次 Laguerre 多项式的零点

$$L_n(x)=e^x\frac{d^n}{dx^n}(e^{-x}x^n)$$

而系数 A_i 为

$$A_i=\frac{1}{L_n'(x_i)}\int_0^\infty\frac{L_n(x)e^{-x}}{x-x_i}dx=\frac{(n!)^2}{x_i[L_n'(x_i)]^2}.$$

截断误差为

$$E=\frac{(n!)^2}{(2n)!}y^{(2n)}(\theta).$$

可以看出,这些结果与 Gauss-Legendre 公式的相应结果非常相似.这里的权函数是 $w(x)=e^{-x}$. n 点公式对直至 $2n-1$ 次的多项式为精确的.自变量值及系数提供在表 15.2 中.

表 15.2

n	x_i	A_i
2	0.58578644	0.85355339
	3.41421356	0.14644661
4	0.32254769	0.60315410
	1.74576110	0.35741869
	4.53662030	0.03888791
	9.39507091	0.00053929
6	0.22284660	0.45896467
	1.18893210	0.41700083
	2.99273633	0.11337338
	5.77514357	0.01039920
	9.83746742	0.00026102
	15.98287398	0.00000090
8	0.17027963	0.36918859
	0.90370178	0.41878678
	2.25108663	0.17579499
	4.26670017	0.03334349
	7.04590540	0.00279454
	10.75851601	0.00009077
	15.74067864	0.00000085
	22.86313174	0.00000000
10	0.13779347	0.30844112
	0.72945455	0.40111993
	1.80834290	0.21806829
	3.40143370	0.06208746
	5.55249614	0.00950152
	8.33015275	0.00075301
	1. 84378584	0.00002826
	16.27925783	0.00000042
10	21.99658581	0.00000000
	29.92069701	0.00000000
12	0.11572212	0.26473137
	0.61175748	0.37775928
	1.51261027	0.24408201
	2.83375134	0.09044922
	4.59922764	0.02010238
	6.84452545	0.00266397
	9.62131684	0.00020323
	13.00605499	0.00000837
	17.11685519	0.00000017
	22.15109038	0.00000000
	28.48796725	0.00000000
	37.09912104	0.00000000
14	0.09974751	0.23181558
	0.52685765	0.35378469
	1.30062912	0.25873461
	2.43080108	0.11548289
	3.93210282	0.03319209
	5.82553622	0.00619287
	8.14024014	0.00073989
	10.91649951	0.00005491
	14.21080501	0.00000241
	18.10489222	0.00000006
	22.72338163	0.00000000
	28.27298172	0.00000000
	35.14944366	0.00000000
	44.36608171	0.00000000

15.36　将 Gauss-Laguerre 公式应用于 e^{-x} 的积分.

解　由于 $L_1(x)=1-x$,我们在 $x_1=1$ 处有一零点.系数 $A_1=1/[L_1'(1)]^2$,它也等于 1.因此一点公式为

$$\int_0^\infty e^{-x}y(x)dx\approx y(1).$$

在这种情况下 $y(x)=1$,而我们得到的精确积分,它也等于 1.这没有什么值得惊讶的,因为 $n=1$ 时对任何次数不大于 1 的多项式为精确的这一点是有保证的.事实上取 $y(x)=ax+b$,这个公式的结果就是

$$\int_0^\infty e^{-x}(ax+b)\mathrm{d}x = y(1) = a+b,$$

它就是准确的.

15.37 应用 Gauss-Laguerre 方法于 $\int_0^\infty e^{-x}\sin x\,\mathrm{d}x$

解 易知这个积分的精确值为 1/2. 正弦函数的光滑性以及它的各阶导数的有界性预示着我们的公式会有好结果. 误差估计 $(n!)^2/(2n)!$ 将 $y^{(2n)}$ 代以它的最大值 1. 当 $n=5$ 时减少到 $\frac{1}{924}$, 表明约有 3 位精度. 将 x_i, A_i 代入 $\sum_{i=1}^{n} A_i\sin x_i$, 得出的实际结果为

n	2	6	10	14
$\sum$	0.43	0.50005	0.5000002	0.50000000

所以我们的误差公式多少有些悲观.

15.38 应用 Gauss-Laguerre 方法于 $\int_1^\infty (e^{-t}/t)\mathrm{d}t$.

解 $y(t)=1/t$ 的不光滑性, 意味着它的 n 阶导数

$$y^{(n)}(t) = (-1)^n n!\,t^{-(n+1)}$$

随 n 快速增长, 使人不敢对近似公式过于相信. 作变量变换 $t=x+1$, 使积分转变为在我们的标准区间上求积:

$$\int_0^\infty e^{-x}\frac{1}{e(x+1)}\mathrm{d}x.$$

而误差公式变成

$$E = \left[\frac{(n!)^2}{(2n)!}\right]\left[\frac{(2n)!}{e(\theta+1)^{2n+1}}\right],$$

它简化成 $(n!)^2/e(\theta+1)^{2n+1}$. 假如我们以 0 来取代 θ 可获得最大导数值, 但这样做肯定是不受鼓励的, 不过还没有其他选择来指定它. 以公式

$$\frac{1}{e}\sum_{i=1}^{n}\frac{A_i}{x_i+1}$$

进行实际计算得出这些结果:

n	2	6	10	14
近似值	0.21	0.21918	0.21937	0.21938

由于准确到 5 位的值为 0.21938, 我们发现完全的悲观主义是没有必要的. 这难以捉摸的量 θ 看来可随 n 而增加. 将实际的和理论的误差作一比较可将 θ 确定为:

n	2	6	10
θ	1.75	3.91	5.95

在本例中函数 $y(x)$ 在 $x=-1$ 处有一奇异点. 即使一个复的奇异点只是靠近积分区间也能造成此处显见的慢收敛. (与题 15.29 进行比较.) 假如我们远离奇异点则收敛就会快一些. 例如, 以同样的方法积分同样的函数, 在 5 到 ∞ 的区间上积分得到结果:

n	2	6	10
近似值	0.001147	0.0011482949	0.0011482954

最后一个值几乎准确到 10 位.

15.39　什么是 Gauss-Hermite 公式?

解　☞ 它们具有形式

$$\int_{-\infty}^{\infty} e^{-x^2} y(x) dx \approx \sum_{i=1}^{n} A_i y(x_i),$$

自变量值 x_i 为 n 次 Hermite 多项式

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}(e^{-x^2})$$

的零点,而系数 A_i 为

$$A_i = \frac{2^{n+1} n! \sqrt{\pi}}{[H_n'(x_i)]^2}.$$

截断误差为

$$E = \frac{n! \sqrt{\pi} y^{(2n)}(\theta)}{2^n (2n)!}.$$

可以看出,这些结果与在 Gauss-Legendre 公式的情况极为相似.这里的权函数为 $w = e^{-x^2}$. n 点公式对直至 $2n-1$ 次的多项式为精确的.自变量值及系数提供在表 15.3 中.

表 15.3

n	x_k	A_k	n	x_k	A_k
2	±0.70710678	0.88622693	10	±2.53273167	0.00134365
				±3.43615912	0.00000764
4	±0.52464762	0.80491409	12	±0.31424038	0.57013524
	−1.65068012	0.08131284		±0.94778839	0.26049231
6	±0.43607741	0.72462960		±1.59768264	0.05160799
	±1.33584907	0.15706732		±2.27950708	0.00390539
	±2.35060497	0.00453001		±3.02063703	0.00008574
				±3.88972490	0.00000027
8	±0.38118699	0.66114701	14	±0.29174551	0.53640591
	±1.15719371	0.20780233		±0.87871379	0.27310561
	±1.98165676	0.01707798		±1.47668273	0.06850553
	±2.93063742	0.00019960		±2.09518326	0.00785005
10	±0.34290133	0.61086263		±2.74847072	0.00035509
	±1.03661083	0.24013861		±3.46265693	0.00000472
	±1.75668365	0.03387439		±4.30444857	0.00000001

15.40　应用 Gauss-Hermite 2 点公式于积分 $\int_{-\infty}^{\infty} e^{-x^2} x^2 dx$.

解　☞ 能够得到一个精确结果,所以我们先计算

$$H_2(x) = e^{x^2} \frac{d^2}{dx^2}(e^{-x^2}) = 4x^2 - 2.$$

该多项式的零点为 $x_k = \pm\sqrt{2}/2$. 系数 A_i 易于由题 15.39 的公式获得,为 $\sqrt{\pi}/2$. 2 点公式因此为

$$\int_{-\infty}^{\infty} e^{-x^2} y(x) dx \approx \frac{\sqrt{\pi}}{2}\left[y\left(\frac{\sqrt{2}}{2}\right) + y\left(-\frac{\sqrt{2}}{2}\right)\right].$$

取 $y(x) = x^2$ 它就成为 $\int_{-\infty}^{\infty} e^{-x^2} x^2 dx = \sqrt{\pi}/2$,这是积分的精确值.

15.41　计算 $\int_{-\infty}^{\infty} e^{-x^2} \sin^2 x dx$ 准确到 6 位.

解　☞ Gauss-Hermite 公式产生的结果为:

n	2	4	6	8	10
近似值	0.748	0.5655	0.560255	0.560202	0.560202

这看来预示有6-位精度，而这个结果实际准确到6位，精确积分值为$\sqrt{\pi}(1-e^{-1})/2$，它准确到8位是0.56020226.

15.42 计算$\int_{-\infty}^{\infty}(e^{-x^2}/\sqrt{1+x^2})dx$准确到3位.

解 ☞ 平方根因子不如前题中的正弦函数那样光滑，所以我们不应期望完全同样快地收敛，而确实也办不到.

n	2	4	6	8	10	12
近似值	0.145	0.151	0.15202	0.15228	0.15236	0.15239

值0.152看来就是所求的.

15.43 什么是 Gauss-Chebyshev 公式?

解 ☞ 取$w(x)=1/\sqrt{1-x^2}$的高斯形式为

$$\int_{-1}^{1}\left[\frac{y(x)}{\sqrt{1-x^2}}\right]dx\approx\frac{\pi}{n}\sum_{i=1}^{n}y(x_i),$$

自变量值x_i为n次 Chebyshev 多项式的零点

$$T_n(x)=\cos(n\arccos x).$$

与其外观相反，它确实是一个n次多项式，其零点为

$$x_i=\cos\left[\frac{(2i-1)\pi}{2n}\right].$$

所有系数A_i均简单地为π/n，截断误差为

$$E=\frac{2\pi y^{(2n)}(\theta)}{2^{(2n)}(2n)!}.$$

15.44 应用$n=1$的 Gauss-Chebyshev 公式来验证熟悉的结果

$$\int_{-1}^{1}\left(\frac{1}{\sqrt{1-x^2}}\right)dx=\pi.$$

解 ☞ 对$n=1$我们得到$T_n(x)=\cos(\arccos x)=x$，由于它只有一个零点，我们的公式退化成$\pi y(0)$. 由于$n=1$的 Gauss 公式对不大于一次的多项式为精确的，已知的积分精确地为$\pi\cdot y(0)=\pi$.

15.45 应用$n=3$公式于$\int_{-1}^{1}(x^4/\sqrt{1-x^2})dx$.

解 ☞ 直接地由定义我们得到$T_3(x)=4x^3-3x$. 所以$x_1=0$, $x_2=\sqrt{3}/2$, $x_3=-\sqrt{3}/2$. Gauss-Chebyshev 公式现产生$(\pi/3)\left(0+\frac{9}{16}+\frac{9}{16}\right)=3\pi/8$，它也是精确的.

补 充 题

15.46 证明$P_n'(x)=xP_{n-1}'(x)+nP_{n-1}(x)$. 如下面那样地开始. 从 Legendre 多项式的定义有

$$P_n'(x)=\frac{1}{2^n n!}\frac{d^n}{dx^n}[n(x^2-1)^{n-1}(2x)].$$

应用乘积的n阶导数定理得到

$$P_n'(x)=\frac{n}{2^n n!}\frac{d}{dx}\left[2x\frac{d^{n-1}}{dx^{n-1}}(x^2-1)^{n-1}+2(n-1)\cdot\frac{d^{n-2}}{dx^{n-2}}(x^2-1)^{n-1}\right]$$

$$=\frac{d}{dx}[xP_{n-1}(x)]-(n-1)P_{n-1}(x).$$

15.47　证明$(1-x^2)P_n''(x)-2xP_n'(x)+n(n+1)P_n(x)=0$，如下。令$z=(x^2-1)^n$，则$z'=2nx(x^2-1)^{n-1}$，使$(x^2-1)z'-2nxz=0$。重复地微分该方程，得到

$$(x^2-1)z^{(2)}-(2n-2)xz'-2nz=0,$$

$$(x^2-1)z^{(3)}-(2n-4)xz^{(2)}-[2n+(2n-2)]z'=0,$$

$$(x^2-1)z^{(4)}-(2n-6)xz^{(3)}-[2n+(2n-2)+(2n-4)]z^{(2)}=0,$$

最后地得到

$$(x^2-1)z^{(n+2)}-(2n-2n-2)xz^{(n+1)}-[2n+(2n-2)+(2n-4)$$
$$+\cdots+(2n-2n)]z^{(n)}=0,$$

它简化为　　$(x^2-1)z^{(n+2)}+2xz^{(n+1)}-n(n+1)z^{(n)}=0.$

由于$P_n(x)=z^{(n)}/2^nn!$，立得所要求的结果.

15.48　微分题15.21的结果并与题15.47进行比较来证明

$$xP_n'(x)-P_{n-1}'(x)=nP_n(x).$$

15.49　用题15.21来证明对所有的n，$P_n(1)=1$，$P_n(-1)=(-1)^n$.

15.50　用题15.46来证明$P_n'(1)=\frac{1}{2}n(n+1)$，$P_n'(-1)=(-1)^{n-1}P_n'(1)$.

15.51　用题15.46来证明

$$P_n^{(k)}(x)=xP_{n-1}^{(k)}(x)+(n+k-1)P_{n-1}^{(k-1)}(x).$$

然后应用和差方法来验证

$$P_n^{(2)}(1)=\frac{(n+2)^{(4)}}{(2\cdot4)},\qquad P_n^{(3)}(1)=\frac{(n+3)^{(6)}}{(2\cdot4\cdot6)}.$$

一般地有　　$P_n^{(k)}(1)=\frac{(n+k)^{(2k)}}{2^kk!}=\frac{(n+k)!}{(n-k)!\,2^kk!}.$

由于Legendre多项式为非偶即奇函数，也可证明

$$P_n^{(k)}(-1)=(-1)^{n-k}P_n^{(k)}(1).$$

15.52　用题15.46及15.48来证明$P_{n+1}'(x)-P_{n-1}'(x)=(2n+1)P_n(x)$.

15.53　$P_n(x)$的首项系数，正如所知为$A_n=(2n)!\ /2^n(n!)^2$.证明它还可以写成

$$A_n=1\cdot\frac{3}{2}\cdot\frac{5}{3}\cdot\frac{7}{4}\cdots\frac{2n-1}{n}=\frac{(1\cdot3\cdot5\cdots(2n-1))}{n!}.$$

15.54　对$n=3$的情况计算Gauss-Legendre自变量值与系数，证明自变量值为$x_k=0,\pm\sqrt{\frac{3}{5}}$，以及对$x_k=0$系数为$\frac{8}{9}$，而对其他$x_k$系数为5/9.

15.55　对于$n=5$验证这些Gauss-Legendre自变量值及系数为：

x_k	A_k
0	0.56888889
±0.53846931	0.47862867
±0.90617985	0.23692689

15.56　应用题15.54的3点Gauss公式于正弦函数的积分：$\int_0^{\pi/2}\sin t\,dt$.这个结果与用7点Simpson公式(题14.17)所得的结果相比较情况如何?

15.57　应用Gauss-Legendre 2点公式($n=2$)于$\int_{-1}^{1}\frac{1}{1+t^2}dt$并将之与精确值$\pi/2\approx1.5708$进行比较.

15.58　利用题15.57中的函数$y(t)=1/(1+t^2)$，将导出$n=2$公式的线性配置多项式和3次密切多项式予以图解.(参见题15.25.)

15.59　验证我们的各种公式接近$\int_0^1x^xdx\approx0.7834$(精确到4位)到什么程度?同时还应用某些自变量等距公式于该积分.哪一种算法工作得最好?哪一种最容易用于手算?哪一种最易于通过编程进行自动计算?

15.60 像在题 15.59 中那样应用不同的方法于 $\int_0^{\pi/2} e^{\sin x} dx = 3.1044$ 并决定出那个算法对自动计算为最好的.

15.61 由题 15.35 中给出的定义计算 Laguerre 多项式直到 $n=5$.

15.62 找出 $L_2(x)$ 的零点并验证在表 15.2 中为 $n=2$ 给出的自变量值和系数.

15.63 用题 15.9 中的方法来证明 $L_n(x)$ 是在下面的意义下对任何低于 n 次的多项式来说都是正交的,即

$$\int_0^{\infty} e^{-x} L_n(x) p(x) dx = 0$$

其中 $p(x)$ 为任何一个这种多项式.

15.64 用题 15.10 和 15.11 的方法来证明

$$\int_0^{\infty} e^{-x} L_n^2(x) dx = (n!)^2.$$

15.65 应用 Gauss-Laguerre 2 点公式来得到这些精确结果:

$$\int_0^{\infty} e^{-x} x^2 dx = 2!, \qquad \int_0^{\infty} e^{-x} x^3 dx = 3!.$$

15.66 对 3 点 Gauss-Laguerre 积分找出精确的自变量值与系数.

15.67 用上题的公式来验证

$$\int_0^{\infty} e^{-x} x^4 dx = 4!, \qquad \int_0^{\infty} e^{-x} x^5 dx = 5!.$$

15.68 应用 $n=6$ 和 $n=8$ 公式于"光滑"积分 $\int_0^{\infty} e^{-x} \cos x dx$.

15.69 应用 $n=6$ 和 $n=8$ 公式于"非光滑"积分 $\int_0^{\infty} e^{-x} \log(1+x) dx$.

15.70 证明 $\int_0^{\infty} e^{-(x+1/x)} dx \approx 0.2797$ 精确到四位.

15.71 由题 15.39 中给出的定义计算 Hermite 多项式直到 $n=5$.

15.72 证明 Gauss-Hermite 1 点公式为 $\int_{-\infty}^{\infty} e^{-x^2} y(x) dx \approx \sqrt{\pi} y(0)$. 它对次数不大于 1 的多项式是精确的. 将它应用于 $y(x)=1$.

15.73 导出 $n=3$ 的 Gauss-Hermite 近似的精确公式. 将它应用于 $y(x)=x^4$ 来得到一个精确结果.

15.74 4 点和 8 点公式复制这个结果有多接近?

$$\int_{-\infty}^{\infty} e^{-x^2} \cos x dx = \sqrt{\pi} e^{-1/4} \approx 1.3804.$$

15.75 4 点和 8 点公式与下面这个结果相合到什么程度?

$$\int_0^{\infty} e^{-x^2-1/x^2} dx = \frac{\sqrt{\pi}}{2e^2} \approx 0.11994.$$

15.76 证明 $\int_{-\infty}^{\infty} [e^{-x^2}/(1+x^2)] dx \approx 1.343$ 准确到 3 位.

15.77 计算 $\int_{-\infty}^{\infty} e^{-x^2} \sqrt{1+x^2} dx$ 准确到 3 位.

15.78 计算 $\int_{-\infty}^{\infty} e^{-x^2} \log(1+x^2) dx$ 准确到 3 位.

15.79 应用 $n=2$ 的 Gauss-Chebyshev 公式于下面积分的精确验证:

$$\int_{-1}^{1} \frac{x^2}{\sqrt{1-x^2}} dx = \frac{\pi}{2}.$$

15.80 求下面的积分 $\int_{-1}^{1} [(\cos x)/\sqrt{1-x^2}] dx$ 准确到 3 位.

15.81 求下面的积分 $\int_{-1}^{1} (\sqrt{1+x^2}/\sqrt{1-x^2}) dx$ 准确到 2 位.

第十六章　奇异积分

盲目地应用前两章的公式是不明智的，所有那些公式均建立在函数 $y(x)$ 可以方便地用一个多项式来逼近的假设上. 如果不能方便地逼近，那么这些公式所产生的结果即使不是完全靠不住也只能是差劲的. 可以自慰的是人们肯定不会像下面那样应用 Simpson 法则:

$$\int_1^2 \frac{dx}{x^2-2} \approx \frac{1}{6}\left[-1+4(4)+\frac{1}{2}\right]=\frac{31}{12}.$$

但是不那么明显的奇异点也许会暂时被疏忽. 把基于多项式的公式用于导数有奇点的函数之尝试也不是很恰当的. 由于多项式拥有无穷尽的光滑导数，所以把它们用于导数有奇点的函数是不理想的，所得到的结果常常是差劲的.

奇异积分过程

存在多种用于处理奇异积分的过程，它们或用于奇异的被积函数或用于无穷限积分. 下列各过程将予以举例说明:

1. **略去奇异性**也许会获得成功. 在某些环境下使得能用愈来愈多的自变量值 x，直到获得满意的结果为止.
2. **级数展开**被积函数的全部或部分，然后逐项地积分，假如收敛得足够快，则它是一种受欢迎的过程.
3. **削减奇异性**也就是把积分分解为奇异和非奇异两部分，奇异部分可用上经典的分析方法，非奇异部分可无忧虑地应用我们的近似积分公式.
4. **变量变换**是分析中最有力的武器之一. 这里也许能把困难的奇异性改变成较容易处理的情况，或者它能把奇异性完全移开.
5. **相对于一个参数作微分**，这涉及把所给的积分归入一族积分并随之通过微分来揭示这族积分的某些基本性质.
6. **Gauss 方法**也能处理某些类型的奇异性，如何参照上一章将予以指明.
7. **渐近级数**也是有关的，但这个过程将在下一章中处理.

题　解

16.1　比较应用 Simpson 法则于 $\sqrt{x}$ 的积分在靠近零与远离零处的结果.

解　☞ 首先我们取区间介于 1 与 1.30 之间，步长为 $h=0.05$，因为我们早些时在题 14.11 中已作过这个计算. Simpson 法则给出一个准确到 5 位的结果. 甚至梯形法则给出的误差也仅为 0.00002. 现将 Simpson 法则用于 0 与 0.30 之间的积分. 此时积分区间长度相同，然而包含了 $\sqrt{x}$ 之导数的一个奇点，我们得到 $\int_0^{0.3}\sqrt{x}\,dx \approx 0.010864$. 由于准确值为 1.0954. 我们的结果还准确不到 3 位，误差大了 100 多倍.

16.2　忽略 $\sqrt{x}$ 在导数中的奇异性并应用 Simpson 法则伴以逐次缩小的区间 h，其影响为何?

解　☞ Polya 曾经证明(*Math. Z*. 1933)对于这种类型的函数(本身连续导数有奇点)，Simpson 法则和其他类似类型的公式都应收敛于精确积分. 计算说明这些结果:

$1/h$	8	32	128	512
$\int_0^1 \sqrt{x}\,dx$	0.663	0.6654	0.66651	0.666646

收敛于 2/3 虽慢但还是像该发生的那样出现了.

16.3 确定忽略奇异性并应用 Simpson 法则于下面积分：$\int_0^1 (1/\sqrt{x})dx = 2$ 的影响.

解 ☞ 此处被积函数本身有一个间断，而且是一个无穷大点，然而 Davis 及 Rabinowitz 曾证明(*SIAM Journal*, 1965)存在收敛性.他们发现 Simpson 法则产生这些结果，这表明了忽略奇异性有时是成功的.

$1/h$	64	128	256	512	1024	2048
近似积分	1.84	1.89	1.92	1.94	1.96	1.97

这个收敛也是慢的，但还是像该发生的那样出现了.在当今的计算速度下，慢收敛可能不足以去否定一种算法.然而，通常有这样的问题，在一个长的计算中舍入误差的影响有多大.对这同一个积分梯形公式取 $h=\frac{1}{4096}$时的结果为 1.98 而应用 Gauss 48 点公式到这个区间的 1/4 上时(总共 192 点)得到的是 1.99.

16.4 确定忽略奇异性并应用 Simpson 法则和 Gauss 法则于下面积分的结果：

$$\int_0^1 \frac{1}{x}\sin\frac{1}{x}dx \approx 0.6347.$$

解 ☞ 此处的被积函数有一个无穷间断而且还是高度振荡的.这种组合预期会造成在数值计算上的困难.Davis 与 Rabinowitz(参看上一题)发现 Simpson 法则是失败的.

$1/h$	64	128	256	512	1024	2048
近似积分	2.31	1.69	-0.60	1.21	0.72	0.32

Gauss48-点公式也不见得好些.所以奇异性并不总是能被忽略的.

16.5 计算奇异积分 $\int_0^1 (e^x/\sqrt{x})dx$ 精确到 3 位.

解 ☞ 直接使用 Taylor 级数导出

$$\int_0^1 (e^x/\sqrt{x})dx = \int_0^1\left(\frac{1}{\sqrt{x}} + x^{1/2} + \frac{1}{2}x^{3/2} + \frac{1}{6}x^{5/2} + \cdots\right)dx$$

$$= 2 + \frac{2}{3} + \frac{1}{5} + \frac{1}{21} + \frac{1}{108} + \frac{1}{660} + \frac{1}{4680} + \frac{1}{37\,800} + \cdots = 2.925$$

在头几项之后级数快速地收敛并且假如需要的话更高的精度也是容易达到的.注意 $1/\sqrt{x}$ 的奇异性在该级数中已作为第一项来处理.(参看下题.)

16.6 应用"削减奇异性"方法于题 16.5 的积分.

解 ☞ 记该积分为 I, 我们有

$$I = \int_0^1 \frac{1}{\sqrt{x}}dx + \int_0^1 \frac{e^x - 1}{\sqrt{x}}dx.$$

第一个积分是初等的而第二个没有奇点，然而，由于$(e^x-1)/\sqrt{x}$在 0 附近其性态与$\sqrt{x}$相似，它确有一个奇点在它的一阶导数中.正像我们在题 16.1 中所见，这就足够使近似积分不精确.

这种削弱的思想可以推广到将奇异性推入一个高阶导数中.例如，我们的积分还可以写成

$$I = \int_0^1 \frac{1+x}{\sqrt{x}}dx + \int_0^1 \frac{e^x - 1 - x}{\sqrt{x}}dx.$$

如果需要的话还可以从指数函数级数中减去更多的其他的项. 这里的第一个积分值为 $\frac{8}{3}$, 而第二个可以用我们的公式来处理, 虽然级数方法看来在这种情况下也是可行的.

16.7　通过一个变量变换来计算题 16.5 的积分.

解　☞ 变量变换(或替换)可能是积分中最有效的手段. 这里我们令 $t=\sqrt{x}$ 并得到 $I = 2\int_0^1 e^{-t^2}dt$, 它没有任何种类的奇异性, 甚至在它的导数中也没有. 这个积分可以用我们的任何一个公式进行计算也可以用级数展开.

16.8　计算 $\int_0^1 (\cos x)(\log x)dx$ 准确到 6 位小数.

解　☞ 这儿采用的过程如问题 16.5 中所用的. 使用 $\cos x$ 的级数, 积分变成

$$\int_0^1\left(1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+\cdots\right)\log x\,dx.$$

使用初等积分

$$\int_0^1 x^i\log x\,dx = \frac{x^{i+1}}{i+1}\left(\log x - \frac{1}{i+1}\right)\Bigg|_0^1 = -\frac{1}{(i+1)^2},$$

积分被换成级数

$$-1+\frac{1}{3^2 2!}-\frac{1}{5^2 4!}+\frac{1}{7^2 6!}-\frac{1}{9^2 8!}+\cdots.$$

它化简为 -0.946083.

16.9　通过一个变量变换来计算 $\int_1^{\infty}\frac{1}{t^2}\sin\frac{1}{t^2}dt$, 它把无穷区间的积分转换成一个有限区间.

解　☞ 令 $x=1/t$. 则该积分就变成了 $\int_0^1 \sin(x^2)dx$, 它可以用不同的近似方法加以计算. 选择 Taylor 级数展开式导出

$$\int_0^1 \sin(x^2)dx = \frac{1}{3}-\frac{1}{42}+\frac{1}{1320}-\frac{1}{75,600}+\cdots,$$

精确到 6 位只有 4 项起作用, 结果为 0.310268.

16.10　证明用在题 16.9 中的变量变换将 $\int_1^{\infty}\frac{\sin t}{t}dt$ 转换成了一个坏的奇异积分, 所以将积分区间转换成有限长度并不总是有用的一步.

证　☞ 取 $x=1/t$ 我们得到出现在题 16.4 中的积分 $\int_0^1 \frac{1}{x}\sin\frac{1}{x}dx$, 在接近 0 时它强烈地振荡, 使得数值积分几乎不可能. 这个问题的积分最好以下章中所讨论的渐近方法加以处理.

16.11　通过在 $\sin x$ 零点之间的直接估算来计算 $\int_1^{\infty}\frac{1}{x^5}\sin\pi x\,dx$, 因此展开成一个交错级数的一部分.

解　☞ 应用 Gauss 8 点公式到每一个相继的区间(1,2), (2,3), 等等, 获得的结果为

区间	积分	区间	积分
(1,2)	−0.117242	(2,3)	0.007321
(3,4)	−0.001285	(4,5)	0.000357
(5,6)	−0.000130	(6,7)	0.000056
(7,8)	−0.000027	(8,9)	0.000014
(9,10)	−0.000008		

总起来为 −0.11094, 它准确到 5 位.

这种对有限长度的区间的直接估算法从其精神实质来说类似于忽略奇异性方法. 上限实际上被一个有限替换值所取代, 在本情况下取 10. 在其后的对积分的贡献就所要求的精度而言可以考虑为零.

16.12　用对参数微分法来计算 $\int_0^\infty e^{-x^2-1/x^2}dx$.

解　这题说明对积分问题还是有另外的处理方法.我们从将这个问题归入到一族类似的问题开始.对于正的 t,令

$$F(t)=\int_0^\infty e^{-x^2-t^2/x^2}dx.$$

由于这个奇异积分的快速收敛允许在积分号下微分,接着我们得到

$$F'(t)=-2t\int_0^\infty \frac{1}{x^2}e^{-x^2-t^2/x^2}dx.$$

现在引进变量变换 $y=t/x$,它使积分得到可喜的简化

$$F'(t)=-2\int_0^\infty e^{-y^2-t^2/y^2}dy=-2F(t).$$

因此 $F(t)=Ce^{-2t}$而常数 C 可以从已知结果求值:

$$F(0)=\int_0^\infty e^{-x^2}dx=\frac{\sqrt{\pi}}{2}.$$

结果为

$$\int_0^\infty e^{-x^2-t^2/x^2}dx=\frac{1}{2}\sqrt{\pi}e^{-2t}.$$

对于 $t=1$ 的特殊情况,其结果为 0.119938 准确到 6 位.

补　充　题

16.13　比较以 $h=\frac{1}{2}$ 应用 Simpson 法则于 $\int_0^1 x\,dx$ 和 $\int_0^1 x\log x\,dx$ 之结果.

16.14　对题 16.13 中的第二个积分使用逐次缩小的 h 区间,并注意到收敛是趋向精确值 $-1/4$.

16.15　用级数展开法对 $\int_0^1(\sin x)/x^{3/2}dx$ 估计到 3 位.

16.16　应用削减奇异性的方法于题 16.15 中的积分,得到一个初等积分和一个直到二阶导数没有奇点的积分.

16.17　忽略题 16.15 中积分的奇异性并应用 Simpson 法则和 Gauss 公式,持续地使用更多的点.这个结果是否朝着题 16.15 中计算所得值收敛?(定义在零处的被积函数如你所希望的.)

16.18　用关于指数函数的级数来估算 $\int_0^1 e^{-x}\log x\,dx$,准确到 3 位.

16.19　通过忽略奇异性并应用 Simpson 法则及 Gauss 公式计算上题的积分.这个结果是否朝着题 16.18 计算所得值收敛?(定义在零处的被积函数如你所希望的)

16.20　利用级数证明

$$-\int_0^1\frac{\log x}{1-x}dx=-\frac{\pi^2}{6},\qquad \int_0^1\frac{\log x}{1+x}dx=-\frac{\pi^2}{12},\qquad \int_0^1\frac{\log x}{1-x^2}dx=-\frac{\pi^2}{8}.$$

16.21　验证 $\int_0^\infty[e^{-x^2}/(1+x^2)]dx=0.6716$ 准确到 4 位.

16.22　验证 $\int_0^\infty e^{-x}\log x\,dx=-0.5772$ 准确到 4 位.

16.23　验证 $\int_0^\infty e^{-x-1/x}dx=0.2797$ 准确到 4 位.

16.24　验证 $\int_0^\infty e^{-x}\sqrt{x}\,dx=0.8862$ 准确到 4 位.

16.25　验证 $\int_0^1(1/\sqrt{-\log x})dx=1.772$ 准确到 4 位.

16.26　验证 $\int_0^{\pi/2}(\sin x)(\log\sin x)\,dx=-0.3069$ 准确到 4 位.

第十七章　和与级数

数与函数作为和的表达式

在应用数学中数与函数作为有限或无限和的表达式被证明为非常有用的.数值分析以各种方法开发此类表达式,包括下面的:

1. **缩短法**(telescoping method)使一个长的和置换成短的成为可能.对计算机来说它具有明显的优点.经典的例子为

$$\frac{1}{1\cdot 2}+\frac{1}{2\cdot 3}+\frac{1}{3\cdot 4}+\cdots+\frac{1}{n(n+1)}$$
$$=\left(1-\frac{1}{2}\right)+\left(\frac{1}{2}-\frac{1}{3}\right)+\cdots+\left(\frac{1}{n}-\frac{1}{n+1}\right)$$
$$=1-\frac{1}{n+1},$$

在此例中可以看出此法的中心思想.每一项为一个差所置换.

2. **快收敛无穷级数**在数值分析中为主要角色之一.典型的例子是关于正弦与余弦函数的级数.对于被表示的函数而言每个这种级数都是一个计算近似值的极好的算法.

3. **加速法**已经为收敛较慢的级数开发出来.如果为了达到所要求的精度必须用太多的项,那么舍入误差以及随着长时间计算而带来的其他麻烦可能妨碍精度的达到.加速方法修改计算的进程,或者换句话说,它们为了使全程的作业缩短而改变算法.

   Euler 变换是一种常用的加速方法.这变换在较早的章节中导出过.它用另一个通常是更快收敛的级数来置换已知级数.

   比较法是另一种加速的手段.其做法基本上与削减奇异性方法相同,它将一个级数分解为一个类似的但是已知的级数和另一个比原来的收敛得更快的级数.

   特殊的方法可以被设计来加速某些函数的级数表达式.对数函数与反正切函数将被用来作为例证.

4. **Bernoulli 多项式**由

$$B_k(x)=\sum_{i=0}^{k}\binom{k}{i}B_{k-i}x^i$$

给出,其中系数 B_i 由

$$B_0=1,\qquad \sum_{i=0}^{k-1}\binom{k}{i}B_i=0$$

所确定,对于 $k=2,3$,等等.Bernoulli 多项式的性质包括以下的:

$$B_i'(x)=iB_{i-1}(x);$$
$$B_i(x+1)-B_i(x)=ix^{i-1};$$
$$\int_0^1 B_i(x)\mathrm{d}x=0,\qquad 当\ i>0;$$
$$B_i(1)=B_i(0),\qquad 当\ i>1.$$

Bernoulli 数 b_i 定义为

$$b_i=(-1)^{i+1}B_{2i},$$

对 $i=1,2$,等等.

整次幂的和与 Bernoulli 多项式以及 Bernoulli 数相关联.两个这种关系式为

$$\sum_{x=1}^{n}x^p=\frac{B_{p+1}(n+1)-B_{p+1}(0)}{p+1}\quad 及\quad \sum_{k=1}^{\infty}\frac{1}{k^{2i}}=\frac{b_i(2\pi)^{2i}}{2(2i)!}.$$

5. **Euler-Maclaurin 公式**可以通过使用 Bernoulli 多项式被谨慎地导出并且得到一个误差估计.它可以作为一种加速方法.Euler 常数

$$C = \lim\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}-\log n\right)$$

可以用 Euler-Maclaurin 公式来估算.6 项就足以产生几乎 10 位小数的精度.

6. **关于 π 的 Wallis 乘积**为

$$\frac{\pi}{2} = \lim \frac{2\cdot 2\cdot 4\cdot 4\cdot 6\cdot 6\cdots 2k\cdot 2k}{1\cdot 3\cdot 3\cdot 5\cdot 5\cdot 7\cdots(2k-1)(2k+1)}.$$

它用来得到关于大阶乘的 Stirling 级数,它具有形式

$$\log \frac{n!\,e^n}{\sqrt{2\pi}n^{n+1/2}} \approx \frac{b_1}{2n} - \frac{b_2}{3\cdot 4n^3} + \frac{b_3}{5\cdot 6n^5} - \cdots + \frac{(-1)^{k+1}b_k}{(2k)(2k-1)n^{2k-1}},$$

b_i 仍为 Bernoulli 数.较简单的阶乘逼近

$$n! \approx \sqrt{2\pi}n^{n+1/2}e^{-n}$$

正是用了一项 Stirling 级数的结果.

7. **渐近级数**可以看作是加速方法的另一形式.虽然通常为发散的,但它们的部分和具有一种有用的性质.经典的情况包含了形如

$$S_n(x) = \sum_{i=0}^{n} \frac{a_i}{x^i}$$

的和,对所有 x 当 n 趋向无穷时它发散,但是当 x 趋向无穷时,有

$$\lim x^n[f(x) - S_n(x)] = 0.$$

用 $S_n(x)$ 作为 $f(x)$ 的逼近时,对于大的自变量 x 其误差可以简单地通过考察级数被略去的第一项非常容易地估得.Stirling 级数是这类渐近级数的一个知名例子.与它大体相同的想法也可以扩展到其他类型的和.

分部积分可将许多平常的积分转变成渐近级数.对于大的 x 来说,它可能是对这些积分进行估算的最好方法.

题　解

缩短法

17.1　对 $\sum_{i=2}^{n}\log\frac{i-1}{i}$ 进行估算.

解　这是另一种缩短和.我们容易得到

$$\sum_{i=2}^{n}\log\frac{i-1}{i} = \sum_{i=2}^{n}[\log(i-1) - \log i] = -\log n.$$

缩短法当然就是像在第 5 章讨论过的差分项求和.和 $\sum y_i$ 可以方便地估算,假如 y_i 可以表示成一个差分,因为这样一来 $\sum_{i=a}^{b} y_i = \sum_{i=a}^{b}\Delta Y_i = Y_{b+1} - Y_a$.

17.2　对幂和 $\sum_{i=1}^{n} i^4$ 进行估算.

解　由于幂可以用阶乘多项式来表示,它反过来可以表示为差(参看第 4 章),任何这种幂和可以被缩短,在本例中

$$\sum_{i=1}^{n} i^4 = \sum_{i=1}^{n}[i^{(1)} + 7i^{(2)} + 6i^{(3)} + i^{(4)}]$$

$$= \sum_{i=1}^{n}\Delta\left[\frac{1}{2}i^{(2)} + \frac{7}{3}i^{(3)} + \frac{6}{4}i^{(4)} + \frac{1}{5}i^{(5)}\right]$$

$$= \frac{1}{2}(n+1)^{(2)} + \frac{7}{3}(n+1)^{(3)} + \frac{6}{4}(n+1)^{(4)} + \frac{1}{5}(n+1)^{(5)}$$

$$= \frac{1}{30}n(n+1)(2n+1)(3n^2+3n-1).$$

其他的幂和可以类似的方式处理.

17.3 对 $\sum_{i=1}^{n}(i^2+3i+2)$ 进行估算.

解 ☞ 由于幂和可以通过和差进行估算,所以多项式值的和是易得的副产品.例如:

$$\sum_{i=1}^{n} i^2 + 3\sum_{i=1}^{n} i + \sum_{i=1}^{n} 2 = \frac{n(n+1)(2n+1)}{6} + \frac{3n(n+1)}{2} + 2n.$$

17.4 对 $\sum_{i=1}^{n}\frac{1}{i(i+1)(i+2)}$ 进行估算.

解 ☞ 它也能够写成差分的和,回想第4章中具有负指数的阶乘多项式,我们得到

$$\frac{1}{2i(i+1)} - \frac{1}{2(i+1)(i+2)} = \frac{1}{i(i+1)(i+2)}.$$

并由此得出已知和缩短为 $\frac{1}{4} - \frac{1}{2(n+1)(n+2)}$.

在本例中这个无穷级数是收敛的,而且

$$\sum_{i=1}^{\infty}\frac{1}{i(i+1)(i+2)} = \frac{1}{4}.$$

17.5 对 $\sum_{i=1}^{n}\frac{3}{i(i+3)}$ 进行估算.

解 ☞ 像这种简单的有理函数(以及在题17.4中)容易求和.这里

$$\sum_{i=1}^{n}\frac{3}{i(i+3)} = \sum_{i=1}^{n}\left(\frac{1}{i} - \frac{1}{i+3}\right) = 1 + \frac{1}{2} + \frac{1}{3} - \frac{1}{n+1} - \frac{1}{n+2} - \frac{1}{n+3},$$

该无穷级数收敛于 $\sum_{i=1}^{\infty}\frac{3}{i(i+3)} = \frac{11}{6}$.

快收敛级数

17.6 $\sin x$ 展成的 Taylor 级数,要取 x 幂的多少项才能对0与 $\pi/2$ 之间的所有自变量提供8位的精度?

解 ☞ 由于级数 $\sin x = \sum_{i=0}^{\infty}(-1)^i x^{2i+1}/(2i+1)!$ 为具有稳定递减项的交错级数.只用 n 项时其截断误差将不超过第 $(n+1)$ 项.这类级数的这个重要性质使截断误差的估计变得相对地容易.这里我们得到 $(\pi/2)^{15}/15! \approx 8\cdot 10^{-10}$,所以正弦级数取7项对整个区间中得到8位精度是确定的.

这是一个快收敛级数的例子.由于其他的自变量可以通过函数的周期性来处理,所以所有的自变量均被覆盖.然而要注意,在自变量的化简中会出现有效数字的严重损失.例如,取 $x \approx 31.4$ 我们得到

$$\sin x \approx \sin 31.4 = \sin(31.4 - 10\pi) \approx \sin(31.4 - 31.416)$$

$$= \sin(-0.016) \approx -0.016$$

用同样的方法 $\sin 31.3 \approx -0.116$ 而 $\sin 31.5 \approx 0.084$.这就意味着,虽然输入数据31.4已知有3位有效数字,可是输出数字连一位有效数字都不肯定.基本上是自变量 x 小数点右边的数字位数决定所得到的 $\sin x$ 的精度.

17.7 e^x 展成 Taylor 级数要取 x 幂的多少项才能对0与1之间的所有自变量提供8位的精度?

解 ☞ 这级数就是熟悉的 $e^x = \sum_{i=0}^{\infty} x^i/i!$.由于它不是交错级数,其截断误差不会是小于弃掉的第一项.这里我们采取一个简单的比较试验.假如我们将级数在 x^n 项后截断,则其误差为

$$\sum_{i=n+1}^{\infty}\frac{x^i}{i!} = \frac{x^{n+1}}{(n+1)!}\left[1 + \frac{x}{n+2} + \frac{x^2}{(n+2)(n+3)} + \cdots\right],$$

且由于 $x<1$ 该误差将不会超过

$$\frac{x^{n+1}}{(n+1)!}\left[1+\frac{1}{n+2}+\frac{1}{(n+2)^2}+\cdots\right]$$
$$=\frac{x^{n+1}}{(n+1)!}\cdot\frac{1}{1-1/(n+2)}$$
$$=\frac{x^{n+1}}{(n+1)!}\cdot\frac{n+2}{n+1}.$$

故它勉强超过第一个略去的项. 当 $n=11$ 时这个误差界约为 $2\cdot10^{9}$. 因而表明多项式为 11 次的. 例如, 在 $x=1$ 处相继的项如下:

1.00000000	0.50000000	0.04166667	0.00138889	0.00002480	0.00000028
1.00000000	0.16666667	0.00833333	0.00019841	0.00000276	0.00000003

它们的总和为 2.71828184, 它在最后一位上由于舍入有一个单位的误差.

这个误差曾经能够用 Lagrange 形式(题 11.4)也作过估计, 它给出

$$E=\frac{1}{(n+1)!}e^{\xi}x^{n+1},\qquad 其中 0<\xi<x.$$

17.8 计算 e^{-10} 到 6 个有效数字.

解 这题作为例子说明一个重要的差别. 对于 6 位我们可以如在题 17.7 中对 $x=-10$ 那样进行. 然而这个级数收敛得很慢且有另一类的麻烦. 在作为二个较大数的差来获得这个小数时我们要丢失有效位. 工作到 8 位来获得 $e^{-10}\approx0.00004540$, 它只有 4 位有效数字. 在交错级数中常有这种丢失. 双精度算术(以两倍多的位数进行工作)偶尔会克服这种困难. 然而, 这里我们简单地计算 e^{10} 然后取其倒数. 其结果是 $e^{-10}\approx0.0000453999$, 它准确到最后一位.

17.9 在题 14.34 中积分 $(2/\sqrt{\pi})\int_0^x e^{-t^2}dt$ 是以 Taylor 级数法在 $x=1$ 处进行计算的. 假设该级数用于大的 x 值, 然而为了避免舍入误差的增长取不超过 20 项相加. 与 4 位精确相容 x 可以取得多大?

解 被积级数的第 n 项符号除外为 $2x^{2n-1}/\sqrt{\pi}(2n-1)(n-1)!$. 由于这个级数是具有稳定递减项的交错级数, 截断误差不会超过第一个被略去的项.

使用 20 项我们要求 $(2/\sqrt{\pi})x^{41}/41\cdot20!<5\cdot10^{-5}$. 它近似地导出 $x<2.5$. 对于这种自变量级数快速地收敛足以满足我们的约定. 对更大的自变量并不如此了.

加速方法

17.10 并不是所有的级数收敛得如前面那些题那样快. 由二项式级数

$$\frac{1}{1+x^2}=1-x^2+x^4-x^6+\cdots.$$

从 0 到 x 积分我们得到

$$\arctan x=x-\frac{1}{3}x^3+\frac{1}{5}x^5-\frac{1}{7}x^7+\cdots,$$

在 $x=1$ 处它给出 Leibnitz 级数

$$\frac{\pi}{4}=1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\cdots,$$

为得到 4-位精度需要这个级数的多少项?

解 由于这是一个具有稳定递减项的交错级数, 所以截断误差不会超过被省略的第一项. 假如这个项为 0.00005 或更小一些, 我们必须用大约为 1/20,000 以外的项, 这也就是要 10,000 项. 对这么大量的项求和我们可以预计舍入误差将积累到最大单位舍入误差的 100 倍. 然而在情况差到难以置信时该累积甚至可能增长到最大误差的 10,000 倍. 无论如何该级数都不可能导出关于计算 $\pi/4$ 的令人满意的算法.

17.11 应用 11 章中的 Euler 变换于上题中的级数以得到 4 位精度.

解 最佳的过程是将前几项加起来并对剩下的部分应用变换. 例如, 到 5 位数

$$1-\frac{1}{3}+\frac{1}{5}-\cdots-\frac{1}{19}\doteq 0.76046$$

列出随后的几个倒数及它们的差分如下：

0.04762				
	414			
0.04348		66		
	348		14	
0.04000		52		3
	−296		−11	
0.03704		41		
	255			
0.03448				

Euler 变换为

$$y_0-y_1+y_2-y_3+\cdots=\sum_{i=0}^{n}\frac{(-1)^i\Delta^i y_0}{2^{i+1}}=\frac{1}{2}y_0-\frac{1}{4}\Delta y_0+\frac{1}{8}\Delta^2 y_0-\cdots,$$

并应用于我们的表得到

$$0.02381+0.00104+0.00008+0.00001=0.02494.$$

最后我们有

$$\frac{\pi}{4}=1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\cdots=0.76046+0.02494=0.78540,$$

它准确到 5 位. 总起来说, 原有级数只要用 15 项而不是 10,000 项. Euler 变换常常会产生像这样的极好的加速, 然而它也有不成功的时候.

17.12　由公式

$$\frac{\pi}{4}=2\arctan\frac{1}{5}+\arctan\frac{1}{7}+2\arctan\frac{1}{8}$$

来计算 $\pi/4$, 进行到 8 位.

解　☞ 这例子说明这函数会有多么特殊的性质可以用来加速收敛. 级数

$$\arctan x=x-\frac{1}{3}x^3+\frac{1}{5}x^5-\frac{1}{7}x^7+\cdots$$

对现在涉及的自变量值都将快速地收敛. 我们发现用不超过该级数的 5 项得到:

$$2\arctan\frac{1}{5}=0.39479112,\qquad \arctan\frac{1}{7}=0.14189705,\qquad 2\arctan\frac{1}{8}=0.24870998,$$

总起来为 0.78539815. 最后的一位应该为 6.

17.13　为估算级数 $\sum_{i=1}^{\infty}\frac{1}{i^2+1}$ 需要取多少项才能准确到三位?

解　☞ 从 $i=45$ 开始的项都比 0.0005 小, 所以这些项中没有一项会独自影响第三位小数. 然而, 由于所有的项都是正的, 显然从 $i=45$ 开始的项的和会影响到第三位, 也许甚至第二位. Stegun 及 Abramowitz(*Journal of SIAM*, 1956)证明对 3 位精度来说实际需要 5745 项. 这是慢收敛正项级数的一个好例子.

17.14　由"比较法"估算题 17.13 的级数准确到 3 位. (这个方法类似于通过削减奇异性来对奇异积分进行估算.)

解　☞ 比较法涉及引进一个具有同样收敛速率的已知级数. 例如

$$\sum_{i=1}^{\infty}\frac{1}{i^2+1}=\sum_{i=1}^{\infty}\frac{1}{i^2}-\sum_{i=1}^{\infty}\frac{1}{i^2(i^2+1)},$$

后面我们要证明右边第一个级数为 $\pi^2/6$. 第二个比其他的收敛更快, 并且我们得到

$$\sum_{i=1}^{\infty}\frac{1}{i^2(i^2+1)}=\frac{1}{2}+\frac{1}{20}+\frac{1}{90}+\frac{1}{272}+\frac{1}{650}+\frac{1}{1332}+\frac{1}{2450}+\cdots\approx 0.56798,$$

正好用了 10 项. 从 $\pi^2/6=1.64493$ 中减去它, 最终结果为 1.07695, 它可以舍入到 1.077.

17.15　证明问题 17.14 中得到的结果至少准确到 3 位.

证 ☞ 我们级数计算的截断误差为

$$E = \sum_{i=11}^{\infty} \frac{1}{i^2(i^2+1)} < \sum_{i=11}^{\infty} \frac{1}{i^4} = \sum_{i=1}^{\infty} \frac{1}{i^4} - \sum_{i=1}^{10} \frac{1}{i^4}.$$

稍后要证明右边第一个级数为 $\pi^4/90$，而第二个至少为 1.08200. 这样 $E < 1.08234 - 1.08200 = 0.00034$. 因为相加的是 11 个有 5 位精度的数，所以其舍入误差不会超过 $11 \cdot 5 \cdot 10^{-6}$. 因此，合起来的误差不会超过 0.0004，使得我们的结果能准确到 3 位.

17.16 应用比较法于 $\sum_{i=1}^{\infty} \frac{1}{i^2(i^2+1)}$.

解 ☞ 这个级数在上题中是直接求和的. 然而，以此为例说明比较法如何反复地被应用，请注意到

$$\sum_{i=1}^{\infty} \frac{1}{i^2(i^2+1)} = \sum_{i=1}^{\infty} \frac{1}{i^4} - \sum_{i=1}^{\infty} \frac{1}{i^4(i^2+1)},$$

直接估算后一积分得出 $\frac{1}{2}+\frac{1}{80}+\frac{1}{810}+\frac{1}{4352}+\frac{1}{16,250}+\cdots$，它得出 0.51403. 从 $\pi^4/90$ 中减去它，我们得到

$$\sum_{i=1}^{\infty} \frac{1}{i^2(i^2+1)} \approx 1.08234 - 0.51403 = 0.56831$$

它与上两题的结果十分一致，那里同样的和计算得 0.56798 带有一个估计误差为 0.00034，该误差估计几乎是完美的.

17.17 估算 $\sum_{i=1}^{\infty} \frac{1}{i^3}$ 到 4 位.

解 ☞ 这个级数收敛得过于慢些而不能令人满意. 应用比较法

$$\sum_{i=1}^{\infty} \frac{1}{i^3} = 1 + \sum_{i=2}^{\infty} \frac{1}{(i-1)i(i+1)} - \sum_{i=2}^{\infty} \frac{1}{i^2(i^3-i)}$$

右端第一个级数是可缩短的，在题 17.4 中已知它精确地为 $\frac{1}{4}$. 后一个可以直接求和，

$$\frac{1}{24}+\frac{1}{216}+\frac{1}{960}+\frac{1}{3000}+\frac{1}{7560}+\frac{1}{16,464}+\cdots$$

得出为 0.04787，从 1.25 中减去它，最后我们得到 $\sum_{i=1}^{\infty} 1/i^3 = 1.20213$，它准确到 4 位，关于一个更精确的结果可参见题 17.39.

Bernoulli 多项式

17.18 Bernoulli 多项式 $B_i(x)$ 定义为

$$e^{xt}\frac{t}{e^t - 1} = \sum_{i=0}^{\infty} \frac{t^i}{i!} B_i(x),$$

令 $B_i(0) = B_i$ 并推出对这些数 B_i 一个递推公式.

解 ☞ 将 x 换成 0，我们有

$$t = (e^t - 1)\sum_{i=0}^{\infty} \frac{t^i B^i}{i!} = \left(\sum_{j=1}^{\infty} \frac{t^j}{j!}\right)\left(\sum_{i=0}^{\infty} \frac{t^i B^i}{i!}\right) = \sum_{k=1}^{\infty} c_k t^k,$$

其中 $c_k = \sum_{i=0}^{k-1} \frac{B_i}{i!(k-i)!}$. 这使得 $k!c_k = \sum_{i=0}^{k-1}\binom{k}{i}B_i$. 比较上面级数方程中 t 的系数，我们发现

$$B_0 = 1, \qquad \sum_{i=0}^{k-1}\binom{k}{i}B_i = 0, \qquad 当\ k = 2, 3, \cdots$$

把它写开来，这个方程组说明 B_i 可以毫无困难地逐个决定：

$$B_0 = 1,$$
$$B_0 + 2B_1 = 0,$$
$$B_0 + 3B_1 + 3B_2 = 0,$$
$$B_0 + 4B_1 + 6B_2 + 4B_3 = 0,$$

等等.因此头几个 B_i 为

$$B_0=1,\quad B_1=-\frac{1}{2},\quad B_2=\frac{1}{6},\quad B_3=0,\quad B_4=-\frac{1}{30},\quad B_5=0,\quad B_6-\frac{1}{42},$$

等等.这所用的方程组也可描述成下面的形式

$$(B+1)^k-B^k=0,\qquad 当\ k=2,3,\cdots.$$

它可以理解为在应用二项式定理之后每个"幂"B^i 可以换成 B_i.

17.19 寻找 Bernoulli 多项式的一个显式公式.

解 ☞ 从定义方程和上面处理过的 $x=0$ 的特殊情况有

$$\left(\sum_{i=0}^{\infty}\frac{x^it^i}{i!}\right)\left(\sum_{j=0}^{\infty}\frac{B_jt^j}{j!}\right)=\sum_{k=0}^{\infty}\frac{t^k}{k!}B_k(x).$$

比较两边 t^k 的系数便得

$$\frac{1}{k!}B_k(x)=\sum_{i=0}^{k}B_{k-i}\frac{1}{i!(k-i)!}x^i$$

或者

$$B_k(x)=\sum_{i=0}^{k}\binom{k}{i}B_{k-i}x^i.$$

头几个 Bernoulli 多项式是

$$B_0(x)=1,\qquad B_3(x)=x^3-\frac{3}{2}x^2+\frac{1}{2}x,$$

$$B_1(x)=x-\frac{1}{2},\qquad B_4(x)=x^4-2x^3+x^3-\frac{1}{30},$$

$$B_2(x)=x^2-x+\frac{1}{6},\qquad B_5(x)=x^5-\frac{5}{2}x^4+\frac{5}{3}x^3-\frac{1}{6}x,$$

等等.公式可以概括成 $B_k(x)=(x+B)^i$,这儿再一次地理解为应用了二项式定理且随后每个"幂"B^i 换成了 B_i.

17.20 证明 $B_i'(x)=iB_{i-1}(x)$.

证 ☞ 定义方程可以写成

$$\frac{te^{xt}}{e^t-1}=1+\sum_{i=1}^{\infty}\frac{t^iB_i(x)}{i!}.$$

对 x 微分并以 t 通除,

$$\frac{te^{xt}}{e^t-1}=\sum_{i=1}^{\infty}\left[\frac{B_i'(x)}{i}\right]\left[\frac{t^{i-1}}{(i-1)!}\right]^*.$$

但是定义方程还可以写成

$$\frac{te^{xt}}{e^t-1}=\sum_{i=1}^{\infty}\left[B_{i-1}(x)\right]\left[\frac{t^{i-1}}{(i-1)!}\right].$$

并比较右侧的系数,$B_i'(x)=iB_{i-1}(x)$ 对于 $i=1,2,\cdots$,同时注意通过对 $B_i(x)=(x+B)^i$ 的形式微分立得同样的结果.

17.21 证明 $B_i(x+1)-B_i(x)=ix^{i-1}$.

证 ☞ 形式上(虽然进行严格的证明不会太困难)从 $(B+1)^k=B^k$ 出发,我们得到

$$\sum_{k=2}^{i}\binom{i}{k}(B+1)^kx^{i-k}=\sum_{k=2}^{i}\binom{i}{k}B^kx^{i-k}$$

或者

$$(B+1+x)^i-i(B+1)x^{i-1}=(B+x)^i-iBx^{i-1}.$$

从 Bernoulli 多项式简要公式(题 17.19)它立刻转化成 $B_i(x+1)-B_i(x)=ix^{i-1}$.

17.22 证明 $B_i(1)=B_i(0)$ 当 $i>1$ 时.

证 ☞ 在上题中将 x 换成 0 立得这个结果.

17.23 证明 $\int_0^1 B_i(x)\mathrm{d}x=0$ 对 $i=1,2,\cdots$.

* 译注:原文 $\frac{t^{i-1}}{(i-1)!}$ 误为 $t^{i-1}(i-1)!$.

证 ☞ 凭借前面的那些题,有

$$\int_0^1 B_i(x)\mathrm{d}x = \frac{B_{i+1}(1) - B_{i+1}(0)}{i+1} = 0.$$

17.24 给出 $B_0(x)=1$,题 17.20 与 17.23 的条件同样也能决定 Bernoulli 多项式.以这种方法决定 $B_1(x)$和 $B_2(x)$.

解 ☞ 由 $B_1'(x)=B_0(x)$得出 $B_1(x)=x+C_1$ 其中 C_1 是一个常数.要 $B_1(x)$的积分为 0, C_1 必须为 $-1/2$.然后由 $B_2'(x)=2B_1(x)=2x-1$ 得出 $B_2(x)=x^2-x+C_2$,要 $B_2(x)$的积分为 0,常数 C_2 必须为 1/6.以这种方法可以依次决定每个 $B_i(x)$.

17.25 证明 $B_{2i-1}=0$ 对 $i=2,3,\cdots$.

证 ☞ 注意

$$f(t) = \frac{t}{e^t-1} + \frac{t}{2} = \frac{t}{2}\cdot\frac{e^t+1}{e^t-1} = B_0 + \sum_{i=2}^{\infty}\frac{B_i t^i}{i!}$$

是一个偶函数,即 $f(t)=f(-t)$.所有 t 的奇次幂的系数必定为零,使得对所有奇数 $i(i=1$ 除外) $B_i=0$.

17.26 定义 Bernoulli 数 b_i.

解 ☞ 这些数定义为 $b_i=(-1)^{i+1}B_{2i}$当 $i=1,2,\cdots$.因此

$$b_1=\frac{1}{6},\qquad b_4=\frac{1}{30},\qquad b_7=\frac{7}{6},$$
$$b_2=\frac{1}{30},\qquad b_5=\frac{5}{66},\qquad b_8=\frac{3617}{510},$$
$$b_3=\frac{1}{42},\qquad b_6=\frac{691}{2730},\qquad b_9=\frac{43,867}{798},$$

这在以题 17.18 中的递推公式来计算相应的数 B_i 是容易验证的.

17.27 以 Bernoulli 多项式对 p 次幂的和进行估算.

解 ☞ 因为由题 17.21 有 $\Delta B_i(x)=B_i(x+1)-B_i(x)=ix^{i-1}$, Bernoulli 多项式提供幂函数的"有限积分".它使缩短幂和成为可能.

$$\sum_{x=0}^{n}x^p = \sum_{x=0}^{n}\frac{1}{p+1}\Delta B_{p+1}(x) = \frac{B_{p+1}(n+1)-B_{p+1}(0)}{p+1}.$$

17.28 用 Bernoulli 数对形如 $\sum_{k=1}^{\infty}1/k^{2i}$ 的和进行估算.

解 ☞ 稍后将证明(参看三角逼近的章节)函数

$$F_n(x) = B_n(x),\qquad 0\leqslant x<1.$$
$$F_n(x\pm m) = F_n(x),\qquad \text{对整数 } m.$$

作为 Bernoulli 函数,具有 1 为周期可以表示为

$$F_n(x) = (-1)^{n/2+1}\cdot n!\frac{2}{(2\pi)^n}\cdot\sum_{k=1}^{\infty}\frac{\cos 2\pi kx}{k^n}.$$

当 n 为偶数时,而当 n 为奇数时它是

$$F_n(x) = (-1)^{(n+1)/2}\cdot n!\frac{2}{(2\pi)^n}\cdot\sum_{k=1}^{\infty}\frac{\sin 2\pi kx}{k^n}.$$

当 n 为偶数时,譬如说 $n=2i$,我们令 $x=0$ 便得到

$$\sum_{k=1}^{\infty}\frac{1}{k^{2i}} = (-1)^{i+1}\frac{F_{2i}(0)(2\pi)^{2i}}{2(2i)!}.$$

但是 $F_{2i}(0)=B_{2i}(0)=B_{2i}=(-1)^{i+1}b_i$,因而 $\sum_{k=1}^{\infty}1/k^{2i} = b_i(2\pi)^{2i}/2(2i)!$.

于特例, $\sum_{k=1}^{\infty}1/k^2 = \pi^2/6$, $\sum_{k=1}^{\infty}1/k^4 = \pi^4/90$ 等等.

17.29 证明 Bernoulli 数皆为正的,并且它们随 i 的增加变得任意地大.

证 ☞ 注意到 $1<\sum_{k=1}^{\infty}1/k^{2i}\leqslant\sum_{k=1}^{\infty}1/k^2=\pi^2/6<2$，我们发现

$$\frac{2(2i)!}{(2\pi)^{2i}}<b_i<\frac{4(2i)!}{(2\pi)^{2i}},$$

特别是所有 b_i 均为正数而且它们随 i 的增加无限止地增长.

17.30 证明当 i 增加时 $\lim\frac{(2\pi)^{2i}}{2(2i)!}b_i=1$.

证 ☞ 它也是由题 17.28 中的级数很快地得到.除 $k=1$ 这项外所有的项当 i 增加时趋于 0 并且因为 $1/x^p$ 是 x 的递减函数.

$$\frac{1}{k^p}<\int_{k-1}^{k}\frac{1}{x^p}\mathrm{d}x,$$

故,当 $p>1$ 时,

$$\sum_{k=2}^{\infty}\frac{1}{k^p}<\int_1^{\infty}\frac{1}{x^p}\mathrm{d}x=\frac{1}{p-1}.$$

当 p 增加时(在我们的情况下 $p=2i$)整个级数有极限为零,确立了所要求的结果.由于这级数的所有项均为正的,还可得出 $b_i>2(2i)!\ /(2\pi)^{2i}$.

Euler-Maclaurin 公式

17.31 用 Bernoulli 多项式导出带有误差估计的 Euler-Maclaurin 公式.(该公式曾在第 11 章中通过算子计算获得过,但是不带误差估计.)

解 ☞ 我们从分部积分开始,并用 $B_1'(t)=B_0(t)=1$ 及 $B_1(1)=-B_1(0)=\frac{1}{2}$的事实.

$$\int_0^1 y(t)\mathrm{d}t=\int_0^1 y(t)B_1'(t)\mathrm{d}t=\frac{1}{2}(y_0+y_1)-\int_0^1 y'(t)B_1(t)\mathrm{d}t.$$

再一次进行分部积分,利用来自题 17.20 的 $B_2'(t)=2B_1(t)$以及 $B_2(1)=B_2(0)=b_1$ 获得

$$\int_0^1 y(t)\mathrm{d}t=\frac{1}{2}(y_0+y_1)-\frac{1}{2}b_1(y_1'-y_0')+\frac{1}{2}\int_0^1 y^{(2)}(t)B_2(t)\mathrm{d}t.$$

下一个分部积分带来的是

$$\frac{1}{2}\int_0^1 y^{(2)}(t)B_2(t)\mathrm{d}t=\frac{1}{6}y^{(2)}(t)B_2(t)\Big|_0^1-\frac{1}{6}\int_0^1 y^{(3)}(t)B_3(t)\mathrm{d}t.$$

但是由于 $B_3(1)=B_3(0)=0$,所以积出的项消失,我们继续进行

$$\frac{1}{2}\int_0^1 y^{(2)}(t)B_2(t)\mathrm{d}t=-\frac{1}{24}y^{(3)}(t)B_4(t)\Big|_0^1+\frac{1}{24}\int_0^1 y^{(4)}(t)B_4(t)\mathrm{d}t$$

$$=\frac{1}{24}b_2(y_1^{(3)}-y_0^{(3)})+\frac{1}{24}\int_0^1 y^{(4)}(t)B_4(t)\mathrm{d}t.$$

因为 $B_4(1)=B_4(0)=B_4=-b_2$.如此进行下去,我们推出结果

$$\int_0^1 y(t)\mathrm{d}t=\frac{1}{2}(y_0+y_1)+\sum_{i=1}^{k}\frac{(-1)^ib_i}{(2i)!}(y_1^{(2i-1)}-y_0^{(2i-1)})+R_k,$$

其中

$$R_k=\frac{1}{(2k)!}\int_0^1 y^{(2k)}(t)B_{2k}(t)\mathrm{d}t.$$

对 R_k 进行分部积分,积出的部分仍为零,导出

$$R_k=\frac{-1}{(2k+1)!}\int_0^1 y^{(2k+1)}(t)B_{2k+1}(t)\mathrm{d}t,$$

相应的结果对其他的在相继整数之间的区间也成立.进行求和,我们发现实质的缩节并得到

$$\sum_{i=0}^{n}y_i=\int_0^n y(t)\mathrm{d}t+\frac{1}{2}(y_0+y_n)-\sum_{i=1}^{k}\frac{(-1)^ib_i}{(2i)!}(y_n^{(2i-1)}-y_0^{(2i-1)}),$$

具有误差

$$E_k=\frac{-1}{(2k+1)!}\int_0^n y^{(2k+1)}(t)F_{2k+1}(t)\mathrm{d}t,$$

其中 $F_{2k}(t)$是题 17.28 中的 Bernoulli 函数,是 Bernoulli 多项式 $B_{2k}(t)$的周期性的拓展.同样的讨论可以用在整数自变量 a 与 b 之间而不是 0 与 n.我们也可以允许 b 变为无穷,假定我们遇到的级数与积分是收敛的.在这种情况下我们假设 $y(t)$和它的导数在无穷远处都变为零,所以公式就变

成了

$$\sum_{i=a}^{\infty} y_i = \int_a^{\infty} y(t)\mathrm{d}t + \frac{1}{2}y_a + \sum_{i=1}^{k}\frac{(-1)^i b_i}{(2i)!}y_a^{(2i-1)}.$$

17.32　通过使用 Euler-Maclaurin 公式对幂和 $\sum_{i=0}^{n} i^4$ 进行估算.

解　在这种情况下函数 $y(t)=t^4$,所以取 $k=2$ 前题的级数有界.此外,E_k 变成零.因为 $y^{(5)}(t)$ 为零.其结果与题 17.2 的相同.

$$\sum_{i=0}^{n} i^4 = \frac{1}{5}n^5 + \frac{1}{2}n^4 + \frac{1}{12}(4n^3) - \frac{1}{720}(24n)$$

$$= \frac{1}{30}n(n+1)(2n+1)(3n^2+3n-1).$$

这是一个例子,在该例中在 Euler-Maclaurin 公式中 k 的增长导出一个有限和.(题 17.27 也可能被应用于这和.)

17.33　在假设收敛的情况下计算 Euler 常数 $C=\lim\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}-\log n\right)$.(同时参看题 17.77)

解　利用题 17.1,它可以被写成 $C = 1 + \sum_{i=2}^{\infty}\left(\frac{1}{i}+\log\frac{i-1}{i}\right)$.

Euler-Maclaurin 公式现可以应用于 $y(t)=1/t-\log t+\log(t-1)$.事实上将头几项直接求和然后在级数的剩余部分应用 Euler-Maclaurin 公式更为方便.计算到 8 位

$$1+\sum_{i=2}^{9}\left(\frac{1}{i}+\log\frac{i-1}{i}\right) = 0.63174368,$$

使用 10 与∞为积分限,我们首先计算

$$\int_{10}^{\infty}\left[\frac{1}{t}-\log t+\log(t-1)\right]\mathrm{d}t = (1-t)\log\frac{t}{t-1}\Big|_{10}^{\infty}$$

$$= -1 + 9\log 10 - 9\log 9 \approx -0.05175536,$$

第一项来自于上限,通过"不定式"计算而得.接着

$$\frac{1}{2}y_{10} = -0.00268026,\qquad -\frac{1}{12}y'_{10} = -0.00009259$$

$$\frac{1}{720}y_{10}^{(3)} = 0.00000020,$$

在无穷远处所有的值均为零.将刚才所得的 5 项相加得 $C\approx 0.57721567$.用 10 位进行计算并只要再多加一项会导出更好的近似 $C\approx 0.5772156650$,它本身只是在第 10 位上大了一个单位.

在此例中用 Euler-Maclaurin 公式可得到的精度是有限的,到一个转折点之后,使用更多的项(增加 k)将导致对 Euler 常数更差的逼近而不是更好.换言之,我们用了一个发散级数的少数几项来得到我们的结果.为了说明它,我们只需注意到该级数的第 i 项,它是 $\frac{(-1)^{i+1}b_i}{(2i)(2i-1)}\left(\frac{2i+9}{10^{2i}}-\frac{1}{9^{2i-1}}\right)$.并且由题 17.29 知 b_i 超过 $2(2i)!\ /(2\pi)^{2i}$,它保证了这一项无限地增长.对 Euler-Maclaurin 级数来说发散较之收敛更为典型.

17.34　一辆卡车在燃料满负荷的情况下行驶的距离为 1"leg".证明,假如在沙漠边缘有一个无限制的燃料供应的话,那么该卡车就可以横跨无论有多宽的沙漠.估计需要多少燃料才能跨过宽为 10"leg"的沙漠.

解　卡车在燃料的一个满载下在沙漠上可以行驶 1leg 宽.有二次满载的话可以采用如下策略:加满油的卡车在沙漠上行驶到 1/3leg 处把所加油的 1/3 留在那里的一个油罐中,卡车返回到沙漠边上的燃料仓库、油正好用完.第二次加满油后驶至油罐处,将存的油再加上,这样一次满载卡车可以由此驶出 1"leg",因而就可以在沙漠上行驶 $1+\frac{1}{3}$leg,如图 17.1 所示.在仓库有 3 个满载的燃料可用时,两个来回能够在深入沙漠 $\frac{1}{5}$leg 处建立一个存放着够 $\frac{6}{5}$ 满载的油罐.第三次加满油后卡车开到油罐处就有 $\left(\frac{4}{5}+\frac{6}{5}\right)$ 满载可用.重复前面的策略所允许达到的旅程就是 $1+\frac{1}{3}+\frac{1}{5}$leg,如

图 17.2 所示.

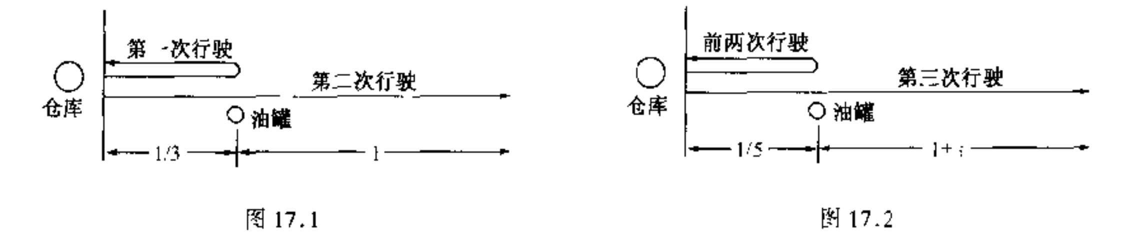


图 17.1　　　　图 17.2

一个类似的策略,允许用 n 次满载的燃料,在沙漠上是 $\left(1+\frac{1}{3}+\frac{1}{5}+\cdots+\frac{1}{2n-1}\right)$ leg 远.由于这个和数当 n 增加时可以任意地增大.所以只要在仓库里有足够多的油可用就能够横跨任意宽的沙漠.

为了估计经过一个 10leg 宽的沙漠所需要的燃料有多少,我们写

$$1+\frac{1}{3}+\cdots+\frac{1}{2n-1}=\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{2n}\right)-\frac{1}{2}\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}\right),$$

并应用题 17.33 的逼近:

$$1+\frac{1}{3}+\cdots+\frac{1}{2n-1}\approx\log(2n)+C-\frac{1}{2}(\log n+C)$$
$$=\frac{1}{2}\log n+\log 2+\frac{1}{2}C\approx\frac{1}{2}\log n+0.98.$$

要它达到 10 对 n 来说几乎是 1 亿次满载燃料.

Wallis 的无穷乘积

17.35　求得 Wallis 关于 π 的乘积.反复地利用从积分表中可以查到的递推公式

解 ☞
$$\int_0^{\pi/2}\sin^n x\,dx=\frac{n-1}{n}\int_0^{\pi/2}\sin^{n-2}x\,dx,\qquad 当\ n>1,$$

容易地带来结果

$$\int_0^{\pi/2}\sin^{2k}x\,dx=\frac{2k-1}{2k}\cdot\frac{2k-3}{2k-2}\cdots\frac{1}{2}\cdot\int_0^{\pi/2}dx,$$
$$\int_0^{\pi/2}\sin^{2k+1}x\,dx=\frac{2k}{2k+1}\cdot\frac{2k-2}{2k-1}\cdots\frac{2}{3}\cdot\int_0^{\pi/2}\sin x\,dx.$$

计算余下的积分并将一个结果除以另一个结果,有

$$\frac{\pi}{2}=\frac{2\cdot2\cdot4\cdot4\cdot6\cdot6\cdots2k\cdot2k}{1\cdot3\cdot3\cdot5\cdot5\cdot7\cdots(2k-1)(2k+1)}\cdot\frac{\int_0^{\pi/2}\sin^{2k}x\,dx}{\int_0^{\pi/2}\sin^{2k+1}x\,dx}.$$

当 k 增加时两个积分的商收敛于 1.它可以证明如下.由于 $0<\sin x<1$,

$$0<\int_0^{\pi/2}\sin^{2k+1}x\,dx\leqslant\int_0^{\pi/2}\sin^{2k}x\,dx\leqslant\int_0^{\pi/2}\sin^{2k-1}x\,dx,$$

用第一个积分去除并用最初的递推公式

$$1\leqslant\frac{\int_0^{\pi/2}\sin^{2k}x\,dx}{\int_0^{\pi/2}\sin^{2k+1}x\,dx}\leqslant\frac{2k+1}{2k},$$

这个商的极限确为 1.因此

$$\frac{\pi}{2}=\lim\frac{2\cdot2\cdot4\cdot4\cdot6\cdot6\cdots2k\cdot2k}{1\cdot3\cdot3\cdot5\cdot5\cdot7\cdots(2k-1)(2k+1)},$$

它就是 Wallis 无穷乘积.

17.36　获得 Wallis 关于 $\sqrt{\pi}$ 乘积.

解 ☞ 由于 $\lim 2k/(2k+1)=1$,上题的结果可以写成

$$\frac{\pi}{2}=\lim\frac{2^2\cdot 4^2\cdots(2k-2)^2}{3^2\cdot 5^2\cdots(2k-1)^2}2k.$$

取平方根并填入缺少的整数，我们得到

$$\sqrt{\frac{\pi}{2}}=\lim\frac{2\cdot 4\cdots(2k-2)}{3\cdot 5\cdots(2k-1)}\sqrt{2k}=\lim\frac{2^{2k}(k!)^2}{(2k)!\sqrt{2k}}.$$

由它立得 Wallis 乘积具有形式

$$\sqrt{\pi}=\lim\frac{2^{2k}(k!)^2}{(2k)!\sqrt{k}},$$

在下题中要用到它.

关于大阶乘的 Stirling 级数

17.37　对大阶乘导出 Stirling 级数.

解　☞ 在 Euler-Maclaurin 公式中令 $y(t)=\log t$ 并用 1 与 n 作为限.于是

$$\log 1+\log 2+\cdots+\log n=n\log n-n+\frac{1}{2}\log n+\sum_{i=1}^{k}\frac{(-1)^i b_i}{(2i)(2i-1)}\left(1-\frac{1}{n^{2i-1}}\right)$$
$$-\int_1^n\frac{F_{2k+1}(t)}{(2k+1)t^{2k+1}}dt.$$

可将它重新安排成

$$\log n!=(n+\frac{1}{2})\log n-n+c-\sum_{i=1}^{k}\frac{(-1)^i b_i}{(2i)(2i-1)n^{2i-1}}$$
$$+\int_n^{\infty}\frac{F_{2k+1}(t)}{(2k+1)t^{2k+1}}dt,$$

其中

$$c=\sum_{i=1}^{k}\frac{(-1)^i b_i}{(2i)(2i-1)}-\int_1^{\infty}\frac{F_{2k+1}(t)}{(2k+1)t^{2k+1}}dt.$$

为了估算 c，在上面的方程中令 $n\to\infty$. 有限和的极限为零. 因为 F_{2k+1} 为周期的所有积分是有界的，其性态与 $1/n^{2k}$ 相同，因而极限也为零. 因此

$$c=\lim\log\frac{n!e^n}{n^{n+1/2}}=\lim\log a_n.$$

现以一个小小的技巧来计算这个极限. 由于 $a_n^2=\frac{(n!)^2e^{2n}}{n^{2n+1}}$，$a_{2n}=\frac{(2n)!\ e^{2n}}{(2n)^{2n+1/2}}$我们发现

$$\lim a_n=\lim\frac{a_n^2}{a_{2n}}=\lim\left[\sqrt{2}\frac{(n!)^2 2^{2n}}{\sqrt{n}(2n!)}\right]=\sqrt{2\pi}.$$

由 Wallis 关于$\sqrt{\pi}$的乘积，因此 $c=\log\sqrt{2\pi}$. 现在我们的结果可以写成 Stirling 级数

$$\log\frac{n!e^n}{\sqrt{2\pi}n^{n+1/2}}\sim\frac{b_1}{2n}-\frac{b_2}{3\cdot 4n^3}+\frac{b_3}{5\cdot 6n^5}-\cdots+\frac{(-1)^{k+1}b_k}{(2k)(2k-1)n^{2k-1}},$$

误差为 $E_n=\int_n^{\infty}\frac{F_{2k+1}(t)}{(2k+1)t^{2k+1}}dt$. 对于大的 n 这意味着对数接近于零，使得 $n!=\sqrt{2\pi}n^{n+1/2}e^{-n}$.

17.38　用 Stirling 级数逼近 20!.

解　☞ 对于 $n=20$ 级数本身就变成了$\frac{1}{240}-\frac{1}{2,880,000}+\cdots\approx 0.00417$ 到 5 位，只用了一项. 现在我们有

$$\log 20!\approx 0.00417-20+\log\sqrt{2\pi}+20.5\log 20\approx 42.33558$$
$$20!\approx 2.43281\cdot 10^{18},$$

它几乎准确到 5 位. 为了更高的精度，Stirling 级数的更多项可能被使用，然而重要的是认识这个级数不是收敛的. 对固定的 n，当 k 增加超过某一点时，项数增加误差 E 增长更快. 这一点从 $b_k>2(2k)!/(2\pi)^{2k}$的事实可以得出(参看题17.29). 正如将被简短地证明，Stirling 级数是渐近级数的一例.

17.39　计算 $\sum_{i=1}^{\infty}1/i^3$ 到 7 位.

解 ☞ 将头 9 项和直接地加起来得到 $\sum_{i=1}^{9} 1/i^3 = 1.19653199$. 用 $f(t) = 1/t^3$ Euler-Maclaurin 公式现在包含

$$\int_{10}^{\infty} \frac{dx}{x^3} = 0.005 \qquad \frac{1}{2} f(10) = 0.0005$$

$$-\frac{1}{12} f'(10) = 0.000025 \qquad \frac{1}{720} f^{(3)}(10) = 0.00000008$$

总起来就是 1.2020569. 它改进了题 17.17 的结果.

渐近级数

17.40 定义一个渐近级数.

解 ☞ 令 $S_n(x) = \sum_{i=0}^{n} a_i x^i$. 如果当 $x \to 0$ 时 $\lim [f(x) - S_n(x)]/x^n = 0$ 对任何固定的正整数 n 成立, 则 $f(x)$ 称作在零这点上渐近于 $\sum_{i=0}^{\infty} a_i x^i$. 它用符号来表示

$$f(x) \approx \sum_{i=0}^{\infty} a_i x^i.$$

用 $x - x_0$ 来代替 x, 应用同样的定义, 级数在 x_0 处渐近于 $f(x)$.

或者所有情况中最有用的是在无穷远处的渐近展式. 如果当 $x \to \infty$ 时

$$\lim x^n [f(x) - S_n(x)] = 0,$$

其中现在 $S_n(x) = \sum_{i=0}^{n} a_i / x^i$, 则 $f(x)$ 在无穷远处有一个渐近级数, 并且我们把它写成

$$f(x) \approx \sum_{i=0}^{\infty} \frac{a_i}{x^i},$$

这个思想可以进一步地推广. 例如, 如果

$$\frac{f(x) - g(x)}{h(x)} \approx \sum_{i=0}^{\infty} \frac{a_i}{x^i},$$

那么我们也可以说 $f(x)$ 有下面的渐近表达式

$$f(x) \approx g(x) + h(x) \sum_{i=0}^{\infty} \frac{a_i}{x^i}.$$

注意这些级数中没有一个被假设为收敛的.

17.41 获得一个关于 $\int_x^{\infty} (e^{-t}/t) dt$ 的渐近级数.

解 ☞ 逐次进行分部积分就带来

$$f(x) = \int_x^{\infty} \frac{e^{-t}}{t} dt = \frac{e^{-x}}{x} - \int_x^{\infty} \frac{e^{-t}}{t^2} dt$$

$$= \frac{e^{-x}}{x} - \frac{e^{-x}}{x^2} + 2! \int_x^{\infty} \frac{e^{-t}}{t^3} dt,$$

等等. 最后人们得到

$$f(x) = \int_x^{\infty} \frac{e^{-t}}{t} dt = e^{-x} \left[\frac{1}{x} - \frac{1}{x^2} + \frac{2!}{x^3} - \frac{3!}{x^4} + \cdots + (-1)^{n+1} \frac{(n-1)!}{x^n} \right] + R_n,$$

其中 $R_n = (-1)^n n! \int_x^{\infty} \frac{e^{-t}}{t^{n+1}} dt$. 由于 $|R_n| < n! e^{-x}/x^{n+1}$, 我们有

$$\left| x^n \left[e^x f(x) - \sum_{i=1}^{n} \frac{(-1)^{i-1} (i-1)!}{x^i} \right] \right| < \frac{n!}{x},$$

所以当 $x \to \infty$ 时它的极限确为零. 这使得 $e^x f(x)$ 渐近于级数而且由我们推广的定义

$$f(x) \approx e^{-x} \left(\frac{1}{x} - \frac{1}{x^2} + \frac{2!}{x^3} - \frac{3!}{x^4} + \cdots \right).$$

注意这个级数对每个 x 值都发散.

17.42 证明在使用上题的级数时所涉及的截断误差不会超过略去的第一项.

证 ☞ 截断误差恰好是 R_n. 略去的第一项为 $(-1)^{n+2} e^{-x} n! / x^{n+1}$, 它等同于在题 17.41 中出

现的 R_n 的估计.

17.43　利用题 17.41 的渐近级数来计算 $f(5)$.

解　☞ 我们发现

$$e^5 f(5) \approx 0.2 - 0.04 + 0.016 - 0.0096 + 0.00746 - 0.00746 + \cdots,$$

在这以后的项就增大.由于误差不超过我们略去的第一项,只需要用 4 项,其结果为

$$f(5) \approx e^{-5}(0.166) \approx 0.00112,$$

最后一位可疑.值得指出的是,该级数不能产生比这更精确的 $f(5)$.对更大的自变量 x,所达到的精度有实质的改进但还是有限的.

17.44　用题 17.41 的级数来计算 $f(10)$.

解　☞ 用 6 位进行计算,我们得到

$$e^{10} f(10) \approx 0.1 - 0.01 + 0.002 - 0.006 + 0.00024 - 0.000120 + 0.000072 - 0.000050 + 0.000040 - 0.000036,$$

在这以后项就增大.将前 9 项相加,我们有 $f(10) \approx e^{-10}(0.091582) \approx 0.0000041579$.最后一位可疑.在上题中可达到 2 位精度.这里我们已运作到 4 位数.渐近级数的基本思想是自变量 x 增大时误差趋于零.

17.45　证明 Stirling 级数是渐近的.

证　☞ 以 n 扮演 x 的角色,$f(x)$ 的角色为对数(参看题 17.37),我们必须证明

$$\lim n^{2k-1} E_n = \lim n^{2k-1} \int_n^\infty \frac{F_{2k+1}(t)}{(2k+1)t^{2k+1}} dt = 0.$$

由于 $F_{2k+1}(t)$ 以周期 1 重复着,故 $F_{2k+1}(t)$* 在区间(0,1)中的性态是有界的,譬如说 $|F| < M$.于是有

$$|n^{2k-1} E_n| < \frac{n^{2k-1} M}{2k(2k+1)n^{2k}},$$

而且随着 n 的增加它变得任意地小.

17.46　找出关于 $\int_x^\infty e^{-t^2/2}$ 的渐近级数.

解　☞ 逐次分部积分法再一次为有效的.首先

$$\int_x^\infty e^{-t^2/2} dt = \int_x^\infty -\frac{1}{t}(-t e^{-t^2/2}) dt = \frac{1}{x} e^{-x^2/2} - \int_x^\infty \frac{1}{t^2} e^{-t^2/2} dt.$$

接着以同样方法继续做下去,我们得到

$$\int_x^\infty e^{-t^2/2} dt = e^{-x^2/2}\left[\frac{1}{x} - \frac{1}{x^3} + \frac{1\cdot 3}{x^5} - \cdots + (-1)^{n-1}\frac{1\cdot 3\cdots(2n-3)}{x^{2n-1}}\right] + R_n,$$

其中 $R_n = 1\cdot 3\cdot 5\cdots(2n-1)\int_x^\infty e^{-t^2/2}\frac{1}{t^{2n}} dt$.这余项可以改写成

$$R_n = \frac{1\cdot 3\cdot 5\cdots(2n-1)}{x^{2n+1}} e^{-x^2/2} - R_{n+1}.$$

因为两个余项均为正的,由此得

$$R_n < \frac{1\cdot 3\cdot 5\cdots(2n-1)}{x^{2n-1}} e^{-x^2/2},$$

它达到了两个目的.一是证明了截断误差不超过第一个省略的项.二是由于它也使 $\lim e^{x^2/2} x^{2n-1} R_n = 0$,证明了该级数为渐近的.

$$\int_x^\infty e^{-t^2/2} dt \approx e^{-x^2/2}\left(\frac{1}{x} - \frac{1}{x^3} + \frac{1\cdot 3}{x^5} - \frac{1\cdot 3\cdot 5}{x^7} + \cdots\right)$$

17.47　以题 17.46 的级数计算 $\sqrt{2/\pi}\int_4^\infty e^{-t^2/2} dt$.

解　☞ 取 $x = 4$ 我们得到

* 译注:原文为 $B_{2k+1}(t)$.

$$\sqrt{\frac{2}{\pi}}e^{-8}[0.25-0.015625+0.002930-0.000916+0.000401-0.000226$$
$$+0.000155-0.000126+0.000118-0.000125+\cdots]$$

一直到项开始增加为止.停止在最小的项之前其结果为

$$\sqrt{\frac{2}{\pi}}\int_4^{\infty}e^{-t^2/2}dt\approx 0.0000633266.$$

2 这个数字可疑.它与我们在题 14.31 中的结果十分相符.独立的计算互相地确认是非常令人放心的.注意到在这二个问题中方法的差异以及目前这个计算的简单性.

17.48 寻求关于正弦积分的一个渐近级数.

解 ☞ 分部积分再一次证明为有用的.首先

$$S_i(x)=\int_x^{\infty}\frac{\sin t}{t}dt=\frac{\cos x}{x}-\int_x^{\infty}\frac{\cos t}{t^2}dt.$$

接着类似地做几步就生成了级数

$$\int_x^{\infty}\frac{\sin t}{t}dt\approx\frac{\cos x}{x}+\frac{\sin x}{x^2}-\frac{2!\cos x}{x^3}-\frac{3!\sin x}{x^4}+\cdots.$$

像上面那些题中那样可以证明它是渐近的.

17.49 计算 $S_i(10)$.

解 ☞ 在上题中令 $x=10$.

$$S_i(10)\approx-0.083908-0.005440+0.001678+0.000326-0.000201-0.000065$$
$$+0.000060+0.000027-0.000034-0.000019$$

在这以后余弦与正弦项都开始增大.这 10 项和舍入到 -0.0876,它准确到 4 位.

补　充　题

17.50 把 $\sum_{i=1}^{n}(i^2-3i+2)$ 表示成差分的和并予以估算.

17.51 把 $\sum_{i=1}^{n}i^5$ 表示成差分的和并予以估算.

17.52 把 $\sum_{i=1}^{n}\frac{1}{i(i+2)}$ 表示成差分的和并予以估算.

17.53 以 Euler-Maclaurin 公式估算题 17.51 的和.

17.54 以 Euler-Maclaurin 公式估算题 17.50 的和.

17.55 需要余弦级数的多少项能在自变量从 0 到 $\pi/2$ 范围内提供 8 位精度?

17.56 证明

$$y_0-y_1+y_2-\cdots=\frac{1}{1+E}y_0=\frac{1}{D}\left(\frac{D}{e^D-1}-\frac{2D}{e^{2D}-1}\right)y_0$$
$$=\left(\frac{1}{2}-B_2\frac{4-1}{2!}D+B_4\frac{16-1}{4!}D^3\right.$$
$$\left.-\frac{64-1}{6!}D^5+\cdots\right)y_0,$$

其中 B_i 为 Bernoulli 数,把它应用于对 $\pi/4$ 的 Leibnitz 级数来获得 6 位结果0.785398.

17.57 应用 Euler 变换估算 $1-\frac{1}{\sqrt{2}}+\frac{1}{\sqrt{3}}-\frac{1}{\sqrt{4}}$到 4 位.

17.58 使用 Euler 变换估算 $1-\frac{1}{9}+\frac{1}{25}-\frac{1}{49}+\cdots$到 8 位,确认这个结果为 0.91596559.

17.59 使用 Euler 变换证明 $1-\frac{1}{\log 2}+\frac{1}{\log 3}-\frac{1}{\log 4}+\cdots$到 4 位等于 0.0757.

17.60 应用 Euler 变换于 $\log 2=1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\frac{1}{5}-\cdots$.

17.61 当自变量 x 值多大时用级数

$$\log(1+x)=x-\frac{1}{2}x^2+\frac{1}{3}x^3-\frac{1}{4}x^4+\cdots$$

的 20 项可以产生有 4 位精度的结果?

17.62 余弦级数 $\cos x=1-\frac{1}{2}x^2+\frac{1}{4!}x^4-\cdots$ 需要多少项才能保证在从 0 到 $\pi/2$ 的区间范围内有 8 位的精度?

17.63 当自变量 x 值多大时用

$$\arctan x=x-\frac{1}{3}x^3+\frac{1}{5}x^5-\frac{1}{7}x^7+\cdots$$

能产生 6 位的精度?

17.64 对级数 $\sinh x=x+\frac{x^3}{3!}+\frac{x^5}{5!}+\frac{x^7}{7!}$ 以被略去的第一项来估计其截断误差.(为找一个可能的方法见题 17.7.)当自变量 x 值多大时用 20 项就足以达到 8 位精度?

17.65 应用题 17.14 的比较法计算 $\sum_{i=1}^{\infty}1/(i^2+i+1)$ 到 3 位.[使用 $\sum_{i=1}^{\infty}1/(i+1)i=1$ 作为比较级数.]

17.66 利用题 17.17 的结果以比较法计算 $\sum_{i=1}^{\infty}1/(i^3+1)$ 到 3 位.

17.67 以比较法计算 $\sum_{i=1}^{\infty}1/(i^2+2i+2)$ 到 3 位.

17.68 以比较法计算 $\sum_{i=1}^{\infty}i^2/(i^4+1)$ 到 3 位.

17.69 从题 17.18 的递推公式确定头 10 个 b_i 数.

17.70 从题 17.19 的公式写出 $B_0(x)$ 到 $B_{10}(x)$.

17.71 证明 $\int_x^{x+1}B_i(x)\mathrm{d}x=x^i$.

17.72 像在题 17.24 中那样确定 $B_3(x)$ 和 $B_4(x)$.

17.73 从 $Q_0(x)=1$ 出发被条件

$$Q_i'(x)=iQ_{i-1}(x),\qquad Q_i(0)=0$$

所确定的是什么样的多项式?

17.74 使用题 17.28 来计算 $\sum_{k=1}^{\infty}1/k^p$ 当 $p=6,8$ 及 10 时的值,验证其结果为 $\pi^6/945,\pi^8/9450$, 及 $\pi^{10}/93\,555$.

17.75 使用 Euler-Maclaurin 公式来证明 $\sum_{i=0}^{n}i^3=n^2(n+1)^2/4$.

17.76 使用 Euler-Maclaurin 公式来对 $\sum_{i=1}^{n}(i^2+3i+2)$ 进行估算.与题 17.3 进行比较.

17.77 使用 Euler-Maclaurin 公式来证明

$$S_n=\sum_{i=1}^{n}\frac{1}{i}-\log n=C+\frac{1}{2n}+\int_n^{\infty}\frac{F_1(t)}{t^2}\mathrm{d}t,$$

其中 C 为 Euler 常数而 $F_1(t)$ 为 $B_1(t)$ 的周期延拓.它证明 S_n 的收敛性并允许当 n 为大数时估计 S_n 及 C 之间的差.

17.78 凭借 Euler-Maclaurin 公式证明

$$C=\frac{1}{2}\log 2+\frac{1}{4}+\sum_{i=1}^{k}\frac{(-1)^{i+1}b_i}{(2i)(2i-1)}\left(\frac{2i+1}{2^{2i}}-1\right)+\text{误差项}$$

并用它估算 Euler 常数 C.证明当 k 增加时,右侧的和变成一个发散级数.从哪一项开始级数项将愈来愈大?

17.79 参照题 17.34,证明横跨一个 5leg 宽的沙漠需要 3000 次以上的满载燃料.

17.80 计算 $\sum_{k=1}^{\infty}1/k^{5/2}$ 到 6 位.

17.81 计算 $\sum_{k=1}^{\infty}1/(2k-1)^2$ 到 3 位.

17.82 精确地估算 $\frac{1}{1}-\frac{1}{4}+\frac{1}{9}-\frac{1}{16}+\frac{1}{25}-\cdots$.

17.83 精确地估算题 17.81 的和.

17.84 证明 Euler 变换将 $\sum_{k=1}^{\infty}\left(-\frac{1}{2}\right)^k$ 转化成一个收敛更快的级数.

17.85　证明 Euler 变换将 $\sum_{k=1}^{\infty}\left(-\frac{1}{3}\right)^k$ 转化成一个收敛更慢的级数

17.86　Stirling 级数产生 2! 具精度为何？在哪一点上级数的项开始增大？

17.87　导出渐近级数

$$\int_x^{\infty}\sin t^2 dt \approx \cos x^2\left(\frac{1}{2x}-\frac{3}{2^3x^5}+\frac{3\cdot 5\cdot 7}{2^5x^9}-\cdots\right)$$

$$+\sin x^2\left(\frac{1}{2^2x^3}-\frac{3\cdot 5}{2^4x^7}+\frac{3\cdot 5\cdot 7\cdot 9}{2^6x^{11}}-\cdots\right)$$

并当 $x=10$ 时使用它，尽你所能得到尽可能高的精度.

第十八章 差分方程

定义

术语差分方程可以期望为指一个包含差分的方程.但是诸如

$$\Delta^2 y_k + 2\Delta y_k + y_k = 0$$

的一个例子它可很快地化简为 $y_{k+2}=0$,这说明差分组合并非永远是方便的,甚至可能是模糊不清的信息.因此差分方程通常直接用 y_k 表出.例如

$$y_{k+1} = a_k y_k + b_k,$$

其中 a_k, b_k 是整数自变量 k 的已知函数.它也可写为 $\Delta y_k=(a_k-1)y_k+b_k$,但它一般并不一定有用.总之,差分方程是定义在一个离散自变量 x_k 的集合上的函数值 y_k 之间的关系式.假定这个自变量等距分布,一般改写自变量 $x_k=x_0+kh$.使我们有了整数自变量 k.

差分方程的解 对某连续整数 k 的集合满足差分方程的一组值 y_k.差分方程的特性使得解序列能递推计算.例如上例中如果已知 y_k 就可以很简单地算出 y_{k+1}.于是某一个已知值就可启动整个序列的计算.

差分方程的阶是差分方程中最大自变量和最小自变量 k 之间的差.上面最后一例是一阶的.

与微分方程的相似之处

差分方程理论和微分方程理论间存在着惊人的相似之处.例如一阶方程一般确有一个解满足初始条件 $y_0=A$.而二阶方程一般确有一解满足两个初始条件 $y_0=A$, $y_1=B$.把这种相似之处进一步强调如下:

1. 在这两个研究课题中**求解的方法**是相似的.对一阶线性差分方程是借助于求和来求解.而相应的一阶微分方程是借助于积分来求解.例如方程 $y_{k+1}=xy_k-c_{k+1}$, $y_0=c_0$ 有多项式解

   $$y_n = c_0x^n + c_1x^{n-1} + \cdots + c_n.$$

   由差分方程本身递推地计算这个多项式来求这个多项式的值.这种方法称为 Horner 方法*.它比通过标准的方幂求值法更经济.

2. **双 Γ 函数**定义如下

   $$\psi(x) = \sum_{i=1}^{\infty} \frac{x}{i(i+x)} - C,$$

   其中 C 是 Euler 常数,它是一阶差分方程

   $$\Delta\psi(x) = \frac{1}{x+1}$$

   的解的求和形式.这也赋予它具有 $1/(x+1)$的有限积分的特征.对于整数自变量 n 可得

   $$\psi(n) = \sum_{k=1}^{n} \frac{1}{k} - C,$$

   这个函数在差分方法中起的作用多少有点类似于微分方程中的对数函数.例如对比这两个公式可得到

---

* 译注:即中国的秦九韶方法.

$$\sum_{k=1}^{\infty} \frac{1}{(k+a)(k+b)} = \frac{\psi(b)-\psi(a)}{b-a},$$

$$\int_1^{\infty} \frac{dx}{(x+a)(x+b)} = \frac{\ln(b+1)-\ln(a+1)}{b-a}.$$

各种和数可以用双 Γ 函数及其导数来表示,上面是一个例子.另一个例子是

$$\sum_{k=1}^{\infty} \frac{2k+1}{k(k+1)^2} = \psi(1) - \psi(0) - \psi'(1),$$

可以证明它等于 $\pi^2/6$.

Γ 函数和双 Γ 函数的关系由下式给出

$$\frac{\Gamma'(x+1)}{\Gamma(x+1)} = \psi(x).$$

3. **二阶线性齐次方程**

$$y_{k+2} + a_1 y_{k+1} + a_2 y_k = 0$$

的解族是
$$y_k = c_1 u_k + c_2 v_k,$$
其中 u_k 和 v_k 本身是方程的解,c_1, c_2 是任意常数,和微分方程理论一样,这称为叠加原理.假设 Wronskian 行列式

$$w_k = \begin{vmatrix} u_k & v_k \\ u_{k-1} & v_{k-1} \end{vmatrix}$$

不等于0,这个方程的任一解都可通过适当选取 c_1 和 c_2 而表为 u_k 和 v_k 的上述形式的叠加.

4. **常系数情形**,a_1 和 a_2 是常数时,可容易地求出 u_k 和 v_k.设 r_1 和 r_2 是特征方程

$$r^2 + a_1 r + a_2 = 0$$

的根.其解是

$u_k = r_1^k$ $\quad v_k = r_2^k$ $\quad$ 当 $a_1^2 > 4a_2$;

$u_k = r^k$ $\quad v_k = kr^k$ $\quad$ 当 $a_1^2 = 4a_2, r_1 = r_2 = r$;

$u_k = R^k \sin k\theta$ $\quad v_k = R^k \cos k\theta$ $\quad$ 当 $a_1^2 < 4a_2, r_1, r_2 = R(\cos\theta \pm i\sin\theta)$.

这与微分方程的相似性是明显的.这些 u_k, v_k 的 Wronskian 行列式不等于0.因此经过叠加我们可求得差分方程的所有解.

***Fibonacci* 数**(斐波那契数)是

$$y_{k+2} = y_{k+1} + y_k$$

的解的值.而且根据上面第一种情形,它能表为实幂函数.它们在信息理论中有某些应用.

5. 非齐次方程
$$y_{k+2} + a_1 y_{k+1} + a_2 y_k = b_k$$
的解族是
$$y_k = c_1 u_k + c_2 v_k + Y_k,$$
其中 u_k, v_k 同前,而 Y_k 是所给方程的一个解.这也类似于微分方程的结果.对某些初等函数 b_k,推导相应的解 Y_k 非常简单.

差分方程的重要性

我们对差分方程的兴趣有两方面.首先它们的确出现在应用中.其次很多求微分方程近似解的方法中包括用差分方程代替微分方程.

题 解

一阶方程

18.1 给定初始条件 $y_0 = 1$.递推地求解一阶方程 $y_{k+1} = ky_k + k^2$.

解 ☞ 这个例子说明了差分方程在计算上的优点.只要进行简单的加法和乘法运算就能逐次求得 y_k 的值,

$$y_1 = 0,\qquad y_2 = 1,\qquad y_3 = 6,\qquad y_4 = 27,\qquad y_5 = 124,$$

等等.差分方程初值问题永远能以这种简单的递推方法求得,但是我们常常希望了解解函数的特性.这就需要求出解的解析表达式,但是仅在某些情形才能求得这样的表达式.

18.2　给定函数 a_k 和 b_k,带有初始条件 $y_0 = A$ 的一阶线性方程 $y_{k+1} = a_k y_k + b_k$ 的解的性质是什么?

解 ☞ 如同上述问题中的方法一样,我们求得

$$y_1 = a_0 A + b_0,$$
$$y_2 = a_1 y_1 + b_1 = a_0 a_1 A + a_1 b_0 + b_1,$$
$$y_3 = a_2 y_2 + b_2 = a_0 a_1 a_2 A + a_1 a_2 b_0 + a_2 b_1 + b_2,$$

等等.以 p_n 表示乘积 $p_n = a_0 a_1 \cdots a_{n-1}$,上述表达式可表现为

$$y_n = p_n\left(A + \frac{b_0}{p_1} + \frac{b_1}{p_2} + \cdots + \frac{b_{n-1}}{p_n}\right),$$

这可以用代入法形式地证明.至于在一阶线性微分方程情形,这个结果仅部分成立.对微分方程而言,解能用积分表示.这里我们有和的形式.然而在某些情形,可以进一步改进.重要的是确实有一个解满足差分方程和预先给定的初始条件 $y_0 = A$.

18.3　在 $a_k = r, b_k = 0$ 的特殊情形解函数的性质是什么?

解 ☞ 这时题 18.2 的结果简化为幂函数 $y_n = Ar^n$.这种幂函数在其他方程的解中也起重要作用.

18.4　当 $a_k = r, b_k = 1$ 和 $y_0 = A = 1$ 时,解函数的性质是什么?

解 ☞ 现在题 18.2 的结果简化为

$$y_n = r^n + r^{n-1} + \cdots + 1 = \frac{r^{n+1} - 1}{r - 1}.$$

18.5　当 $y_0 = A = c_0$ 时,方程 $y_{k+1} = xy_k + c_{k+1}$ 的解函数的性质是什么?

解 ☞ 这一问题很好地说明了有时由差分方程方法可很好地计算简单函数的值.这里题 18.2 的结果变成 $y_n = c_0 x^n + c_1 x^{n-1} + \cdots + c_n$,这个解取为多项式形式.Horner 方法计算这个多项式在自变量 x 处的值需要逐次计算 $y_1, y_2, \cdots, y_n$.这总共有 n 次乘法和 n 次加法.而它等价于把多项式重新写成

$$y_n = c_n + x(c_{n-1} + \cdots + x(c_3 + x(c_2 + x(c_1 + xc_0)))).$$

它比一一地构造 x 的幂,然后用标准多项式计算值更有效.

18.6　当初值 $y_0 = 1$ 时方程 $y_{k+1} = \dfrac{k+1}{x} y_k + 1$ 的解的性质是什么?

解 ☞ 题 18.2 中的 p_n 在这里变为 $p_n = n!\,/x^n$.而所有的 $b_k = 1$.于是解表为

$$\frac{y_n}{p_n} = \frac{x^n y_n}{n!} = 1 + x + \frac{1}{2}x^2 + \cdots + \frac{1}{n!}x^n.$$

因此当 n 增加时,$\lim x^n y_n / n! = e^x$.

18.7　当 $y_0 = 1$ 时,$y_{k+1} = [1 - x^2/(k+1)^2]\,y_k$ 的解的性质是什么?

解 ☞ 题 18.2 中的 b_k 在这里等于 0,而 $A = 1$,使得

$$y_n = p_n = (1 - x^2)\left(1 - \frac{x^2}{2^2}\right)\left(1 - \frac{x^2}{3^2}\right)\cdots\left(1 - \frac{x^2}{n^2}\right).$$

当 $x = \pm 1, \pm 2, \cdots, \pm n$ 时,这个乘积等于零.当 n 增加时我们遇到这个无穷乘积

$$\lim y_n = \prod_{k=0}^{\infty}\left[1 - \frac{x^2}{(k+1)^2}\right],$$

可证明它等于 $(\sin \pi x)/\pi x$.

双 Γ 函数

18.8　"叠进"求和方法,取决于能把一个和数表成差分的和数.

$$\sum_{k=0}^{n} b_k = \sum_{k=0}^{n} \Delta y_k = y_{n+1} - y_0,$$

即需要求解一阶差分方程

$$\Delta y_k = y_{k+1} - y_k = b_k.$$

当 $b_k = 1/(k+1)$时，用这个方法解差分方程并求和.

解 ☞ 由定义双 Γ 函数为 $\psi(x) = \sum_{i=1}^{\infty} \frac{x}{i(i+x)} - C$ 开始，其中 C 是 Euler 常数. 对一切 $x \neq -i$ 可直接求得

$$\Delta\psi(x) = \psi(x+1) - \psi(x) = \sum_{i=1}^{\infty}\left[\frac{x+1}{i(i+x+1)} - \frac{x}{i(i+x)}\right]$$

$$= \sum_{i=1}^{\infty}\left(\frac{1}{i+x} - \frac{1}{i+x+1}\right) = \frac{1}{x+1}.$$

当 x 取整数值时，譬如 $x = k$，这样就提供了求整数倒数和的新的形式，由于

$$\sum_{k=0}^{n-1} \frac{1}{k+1} = \sum_{k=0}^{n-1} \Delta\psi(k) = \psi(n) - \psi(0) = \psi(n) + C,$$

我们也可把它重新写成

$$\psi(n) = \sum_{k=1}^{n} \frac{1}{k} - C.$$

因此对整数自变量，双 Γ 函数是一个熟悉的量. 其性态在图 18.1 表出. 当我们想起 Euler 常数的定义及大的正数 x 的对数的性质，就不会惊奇. 在某种意义上，由 $\psi(n)$引出的 $\psi(x)$很像由 Γ 函数引出的阶乘.

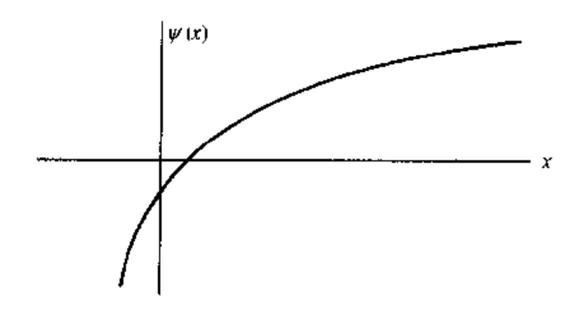


图 18.1

18.9 对任意 t，求和数 $\sum_{k=1}^{n} 1/(k+t)$ 的值.

解 ☞ 根据题 18.8，对任何 x，$\psi(x+1) - \psi(x) = 1/(x+1)$，用 $k+t-1$ 代替 x 得

$$\psi(k+t) - \psi(k+t-1) = \frac{1}{k+t}.$$

现在我们就有了依次叠进和的组成部分，并得到

$$\sum_{k=1}^{n} \frac{1}{k+t} = \sum_{k=1}^{n} [\psi(k+t) - \psi(k+t-1)] = \psi(n+t) - \psi(t).$$

18.10 借助双 Γ 函数求级数 $\sum_{k=1}^{\infty} 1/(k+a)(k+b)$ 的值.

解 ☞ 利用部分分式法得到

$$s_n = \sum_{k=1}^{n} \frac{1}{(k+a)(k+b)} = \frac{1}{b-a}\sum_{k=1}^{n}\left(\frac{1}{k+a} - \frac{1}{k+b}\right).$$

现在利用上题，此式成为

$$s_n = \frac{1}{b-a}[\psi(n+a) - \psi(a) - \psi(n+b) + \psi(b)].$$

根据题 18.8 中的定义，经简单的运算可得

$$\psi(n+a) - \psi(n+b) = (a-b)\sum_{i=1}^{\infty} \frac{1}{(i+n+a)(i+n+b)}.$$

所以当 $n\to\infty$ 时,这个差以零为极限.最后可得

$$\sum_{k=1}^{\infty}\frac{1}{(k+a)(k+b)}=\lim s_n=\frac{\psi(b)-\psi(a)}{b-a}.$$

18.11 求 $\psi'(x),\psi^{(2)}(x)$ 等级数形式的公式.

解 对题 18.8 中的级数微分可得 $\psi'(x)=\sum_{k=1}^{\infty}1/(k+x)^2$.因在任何不含负整数的区间这个级数关于 x 一致收敛,该计算有效.重复之可得

$$\psi^{(2)}(x)=\sum_{k=1}^{\infty}\frac{-2!}{(k+x)^3},\qquad \psi^{(3)}(x)=\sum_{k=1}^{\infty}\frac{3!}{(k+x)^4},\quad 等等.$$

特别地对整数自变量,根据题 17.28 可得 $\psi'(0)=\sum_{k=1}^{\infty}1/k^2=\pi^2/6$.然后我们每次减去一项可得

$$\psi'(1)=\frac{\pi^2}{6}-1,\ \psi'(2)=\frac{\pi^2}{6}-1-\frac{1}{4},一般地\ \psi'(n)=\frac{\pi^2}{6}-1-\frac{1}{4}-\cdots-\frac{1}{n^2}.$$

18.12 求级数 $\sum_{k=1}^{\infty}\frac{2k+1}{k(k+1)^2}$.

解 这进一步说明包含关于 k 的有理项的和或级数是如何通过双 Γ 函数求值的.再次引入部分分式

$$\sum_{k=1}^{\infty}\frac{2k+1}{k(k+1)^2}=\sum_{k=1}^{\infty}\left[\frac{1}{k}-\frac{1}{k+1}+\frac{1}{(k+1)^2}\right].$$

前两项不能分别处理,这是因为其级数是发散的.然而如同在题 18.10 中一样,它们可以在一起处理.其结果是

$$\sum_{k=1}^{\infty}\left[\frac{1}{k(k+1)}+\frac{1}{(k+1)^2}\right]=\psi(1)-\psi(0)+\psi'(1)=\frac{\pi^2}{6}.$$

其他的有理项求和可以类似地处理.

18.13 求级数 $\sum_{k=1}^{\infty}\frac{1}{1^2+2^2+\cdots+k^2}$ 的值.

解 如同在题 5.2 中求平方和,我们可用下式代替此和

$$\sum_{k=1}^{\infty}\frac{6}{k(k+1)(2k+1)}=\sum_{k=1}^{\infty}\left(\frac{6}{k}+\frac{6}{k+1}-\frac{24}{2k+1}\right).$$

因为这三个级数中没有一个自身是收敛的,我们并不分开来处理.推广用于刚才解决问题的方法,我们把上式重写为

$$\begin{aligned}&\sum_{k=1}^{\infty}\left[\left(\frac{6}{k}-\frac{6}{k}\right)+\left(\frac{6}{k+1}-\frac{6}{k}\right)-\left(\frac{24}{2k+1}-\frac{24}{2k}\right)\right]\\&=\sum_{k=1}^{\infty}\left[\frac{-6}{k(k+1)}+\frac{6}{k\left(k+\frac{1}{2}\right)}\right]\\&=-6[\psi(1)-\psi(0)]+12\left[\psi\left(\frac{1}{2}\right)-\psi(0)\right].\end{aligned}$$

这里,题 18.10 在最后一步被用了两次.最后

$$\sum_{k=1}^{\infty}\frac{1}{1^2+2^2+\cdots+k^2}=12\psi\left(\frac{1}{2}\right)-6+12C.$$

18.14 证明 $\mathscr{S}(x)=\Gamma'(x+1)/\Gamma(x+1)$ 也有性质 $\Delta\mathscr{S}(x)=1/(x+1)$,这里 $\Gamma(x)$ 是 Γ 函数.

证 对于正数 x,Γ 函数的定义如下

$$\Gamma(x)=\int_0^{\infty}e^{-t}t^{x-1}dt.$$

分部积分得熟悉的性质　$\Gamma(x+1)=x\Gamma(x)$.

然后在两边求导数得到

$$\Gamma'(x+1)=x\Gamma'(x)+\Gamma(x)$$

或者

$$\frac{\Gamma'(x+1)}{\Gamma(x+1)}-\frac{\Gamma'(x)}{\Gamma(x)}=\frac{1}{x},$$

由此只要用 $x+1$ 代替 x 即得到所需的结论.

因为 $\psi(x+1)\quad\psi(x)=1/(x+1)$,我们得到

$$\frac{\Gamma'(x+1)}{\Gamma(x+1)}-\psi(x)=A,$$

这里 A 是常数,而 x 限制在间隔为 1 的离散集合.对一切除了非负整数外的 x,能证明同样的结论,常数 A 等于 0.

二阶线性方程,齐次情形

18.15　差分方程 $y_{k+2}+a_1y_{k+1}+a_2y_k=0$ 称为线性齐次的,其中 a_1 和 a_2 可能依赖于 k.如果 u_k 和 v_k 是解.那么对任何常数 c_1 和 c_2 证明 $c_1u_k+c_2v_k$ 也是解.(这一性质等同于线性齐次方程.由于 $y_k=0$ 是解,所以这个方程是齐次的.)

证　☞ 因为 $u_{k+2}+a_1u_{k+1}+a_2u_k=0$ 和 $v_{k+2}+a_1v_{k+1}+a_2v_k=0$.第一个方程乘 c_1,第二个方程乘 c_2 并相加得到

$$c_1u_{k+2}+c_2v_{k+2}+a_1(c_1u_{k+1}+c_2v_{k+1})+a_2(c_1u_k+c_2v_k)=0.$$

此即所证.

18.16　对于常数 a_1 和 a_2,证明两个实数解可通过初等函数求得.

证　☞ 首先设 $a_1^2>4a_2$,再取

$$u_k=r_1^k,\qquad v_k=r_2^k,$$

其中 r_1 和 r_2 是二次方程 $r^2+a_1r+a_2=0$ 的相异实根.我们直接可证

$$u_{k+2}+a_1u_{k+1}+a_2u_k=r^k(r^2+a_1r+a_2)=0,$$

这里 r 是二次方程的任一根.此处的二次方程称为特征方程.

其次设 $a_1^2=4a_2$.则特征方程只有一个根,譬如说是 r,特征方程可写为

$$r^2+a_1r+a_2=\left(r+\frac{1}{2}a_1\right)^2=0.$$

现在两个实数解是

$$u_k=r^k,\qquad v_k=kr^k,$$

对于解 u_k 可完全同上面一样得证.至于 v_k

$$(k+2)r^{k+2}+a_1(k+1)r^{k+1}+a_2kr^k$$
$$=r^2[k(r^2+a_1r+a_2)+(2r+a_1)r]=0,$$

这是因为两个括号内都等于 0.

最后设 $a_1^2<4a_2$,则特征方程有一对共轭复根 $Re^{\pm i\theta}$.代入方程得

$$R^2e^{\pm i2\theta}+a_1Re^{\pm i\theta}+a_2$$
$$=R^2(\cos2\theta\pm i\sin2\theta)+a_1R(\cos\theta\pm i\sin\theta)+a_2$$
$$=(R^2\cos2\theta+a_1R\cos\theta+a_2)\pm i(R^2\sin2\theta+a_1R\sin\theta)=0,$$

这要求两个括号内都等于 0.

$$R^2\cos2\theta+a_1R\cos\theta+a_2=0,\qquad R^2\sin2\theta+a_1R\sin\theta=0.$$

现在来证明差分方程的两个实根是

$$u_k=R^k\sin k\theta\qquad v_k=R^k\cos k\theta$$

例如对 u_k 有

$$u_{k+2}+a_1u_{k+1}+a_2$$
$$=R^{k+2}\sin(k+2)\theta+a_1R^{k+1}\sin(k+1)\theta+a_2R^k\sin k\theta$$
$$=R^k(\sin k\theta)(R^2\cos2\theta-a_1R\cos\theta+a_2)+R^k(\cos k\theta)(R^2\sin2\theta+a_1R\sin\theta)$$
$$=0,$$

这是因为上式中的两个括号内都等于 0.对于 v_k 的证明几乎完全相同.

现在可知:对任意常数 a_1 和 a_2,方程 $y_{k+2}+a_1y_{k+1}+a_2=0$ 都有一族基本解 $y_k=c_1u_k+c_2v_k$.

18.17 设 $A>1$.用幂函数求解差分方程 $y_{k+2}-2Ay_{k+1}+y_k=0$.

解 ☞ 设 $y_k=r^k$ 代入方程必须满足 $r^2-2Ar+1=0$.于是得到 $r=A\pm\sqrt{A^2-1}=r_1,r_2$ 并且 $y_k=c_1r_1^k+c_2r_2^k=c_1u_k+c_2v_k$.由于 $r_1>1,0<r_2<1$,[因为 $(A-1)^2=A^2+1-2A<A^2-1$,所以 $r_2=A-\sqrt{A^2-1}<1$.]当 k 变大时,其中一个幂函数成为任意大,而另一个趋于零.

18.18 解方程 $y_{k+2}-2y_{k+1}+y_k=0$.

解 ☞ 这里我们有 $a_1^2=4a_2=4$, $r^2-2r+1=0$.仅有一个根 $r=1$.这表明 $u_k=1$, $v_k=k$ 是方程的解.而 $y_k=c_1+c_2k$ 是方程的解族.根据这个差分方程可以写成 $\Delta^2y_k=0$ 的事实,这一点就毫不奇怪了.

18.19 解 $y_{k+2}-2Ay_{k+1}+y_k=0$,这里 $A<1$.

解 ☞ 现在 $a_1^2<4a_2$,特征方程的根变成

$$Re^{\pm i\theta}=A\pm i\sqrt{1-A^2}=\cos\theta\pm i\sin\theta,$$

其中 $A=\cos\theta$, $R=1$.因此 $u_k=\sin k\theta$, $v_k=\cos k\theta$.于是有解族

$$y_k=c_1\sin k\theta+c_2\cos k\theta.$$

函数 v_k 表示为 A 的多项式时,就被称为 Chebyshev 多项式.例如

$$v_0=1\qquad v_1=A\qquad v_2=2A^2-1$$

这一问题的差分方程就是 Chebyshev 多项式的递推公式.

18.20 证明如果 $y_{k+2}+a_1y_{k+1}+a_2y_k=0$ 的两个解在相邻整数 k 取等值,则对一切整数 k,它们都相等(设 $a_2\neq0$).

证 ☞ 设 u_k 和 v_k 是在 $k=m$ 和 $k=m+1$ 时取等值的两个解.于是其差 $d_k=u_k-v_k$ 是解(根据题 18.15).这时 $d_m=d_{m+1}=0$,于是

$$d_{m+2}+a_1d_{m+1}+a_2d_m=0,\quad d_{m+1}+a_1d_m+a_2d_{m-1}=0.$$

由此可推出 $d_{m+2}=0$ 和 $d_{m-1}=0$.同法可证对于 $k>m+2$ 和 $k<m-1$,依次取每一整数 k, d_k 都等于零.于是 d_k 恒等于零,即 $u_k\equiv v_k$.(假设 $a_2\neq0$,只是为了保证确为二阶差分方程.)

18.21 证明 $y_{k+2}+a_1y_{k+1}+a_2y_k=0$ 的任一解可表为两个特解 u_k 和 v_k 的组合.

$$y_k=c_1u_k+c_2v_k$$

只要它们的 Wronsky 行列式

$$w_k=\begin{vmatrix}u_k & v_k\\ u_{k-1} & v_{k-1}\end{vmatrix}\neq0$$

证 ☞ 我们知道 $c_1u_k+c_2v_k$ 是一个解.由上一问题知,如果在 k 的两个相邻整数值上它与 y_k 相等,它就与 y_k 恒等.为此,我们选取 $k=0$ 和 $k=1$(任何其他相邻整数都行).由以下方程来确定系数 c_1 和 c_2

$$c_1u_0+c_2v_0=y_0,\qquad c_1u_1+c_2v_1=y_1.$$

因为 $w_1\neq0$,其惟一解为 $c_1=(y_1v_0-y_0v_1)/w_1$, $c_2=(y_0u_1-y_1u_0)/w_1$.

18.22 证明如果对某一 k 值,Wronsky 行列式等于零,则它必恒等于零,假设 u_k 和 v_k 是题 18.20 中方程的解.将它用于题 18.16 的特殊情形来证明 $w_k\neq0$.

证 ☞ 我们计算差分

$$\begin{aligned}\Delta w_k&=(u_{k+1}v_k-v_{k+1}u_k)-(u_kv_{k-1}-v_ku_{k-1})\\&=v_k(-a_1u_k-a_2u_{k-1})-u_k(-a_1v_k-a_2v_{k-1})-u_kv_{k-1}+v_ku_{k-1}\\&=(a_2-1)w_k=w_{k+1}-w_k\end{aligned}$$

由此,立刻得到 $w_k=a_2^kw_0$ 因为 $a_2\neq0$,所以使 w_k 等于零的惟一方法是必须$w_0=0$,因此 w_k 就恒等于零.

当 w_k 恒等于零时,可得到 u_k/v_k 等于u_{k-1}/v_{k-1}对一切 k 都成立.即 $u_k/v_k=$常数.这对题18.16 中的 u_k 和 v_k 是绝对不可能的.因此那里的 w_k 不可能为零.

18.23 通过直接计算解初值问题

$$y_{k+2}=y_{k+1}+y_k \qquad y_0=0, \quad y_1=1.$$

解 ☞ 取 $k=0,1,2,\cdots$,我们易求得逐次的 y_k 值 1,2,3,5,8,13,21,34,55,89,144,…,它称为 Fibonacci 数.计算清楚地说明这是一个递增的解序列,但并没有给出它的精确的特性.

18.24 确定上题中解的特性.

解 ☞ 沿着在题 18.15,18.16 等等中拟定的历史轨迹,我们考虑特征方程 $r^2-r-1=0$.因为 $a_1^2>4a_2$,所以有两个实根,即 $r_1,r_2=(1\pm\sqrt{5})/2$.一切解都能表为以下形式

$$y_k=c_1u_k+c_2v_k=c_1\left(\frac{1+\sqrt{5}}{2}\right)^k+c_2\left(\frac{1-\sqrt{5}}{2}\right)^k.$$

为了满足初始条件,我们要求 $c_1+c_2=0$ 和 $c_1\left(\frac{1+\sqrt{5}}{2}\right)+c_2\left(\frac{1-\sqrt{5}}{2}\right)=1$.因此 $c_1=-c_2=\frac{1}{\sqrt{5}}$,$y_k=\frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^k-\left(\frac{1-\sqrt{5}}{2}\right)^k\right]$.

18.25 对 Fibonacci 数证明 $\lim(y_{k+1}/y_k)=(1+\sqrt{5})/2$.

证 ☞ 为了得到这一结果,了解函数的性质是方便的.利用前一题,简单计算可得

$$\frac{y_{k+1}}{y_k}=\frac{1+\sqrt{5}}{2}\cdot\frac{1-[(1-\sqrt{5})/(1+\sqrt{5})]^{k+1}}{1-[(1-\sqrt{5})/(1+\sqrt{5})]^k}.$$

而 $(1-\sqrt{5})/(1+\sqrt{5})$ 的绝对值小于 1.于是就得到所求结果.

18.26 Fibonacci 数出现在某些包含信息沿着通讯线路分布的问题中,系统的能力 C 定义为 $C=\lim(\log y_k)/k$,对数以 2 为底.求这个极限的值.

解 ☞ 再一次要用到解 y_k 的解析性质.这一点已经有了.而且我们有

$$\log y_k=\log\frac{1}{\sqrt{5}}+\log\left[\left(\frac{1+\sqrt{5}}{2}\right)^k-\left(\frac{1-\sqrt{5}}{2}\right)^k\right]$$

$$=\log\frac{1}{\sqrt{5}}+\log\left(\frac{1+\sqrt{5}}{2}\right)^k+\log\left[1-\left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^k\right].$$

使得 $C=\lim\left\{\frac{\log(1/\sqrt{5})}{k}+\log\frac{1+\sqrt{5}}{2}+\frac{1}{k}\log\left[1-\left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^k\right]\right\}=\log\frac{1+\sqrt{5}}{2}$.

非齐次情形

18.27 方程 $y_{k+2}+a_1y_{k+1}+a_2y_k=b_k$ 是线性非齐次的.证明如果 u_k 和 v_k 是相应的齐次方程(即 $b_k=0$)的解,其 Wronsky 行列式不为零.若 Y_k 是上述非线性方程的一个特解,则其每一个解都可表为 $y_k=c_1u_k+c_2v_k+Y_k$.这里 c_1 和 c_2 是适当的常数.

证 ☞ 用 y_k 表示非齐次方程的任一解,Y_k 表示其特解,

$$y_{k+2}+a_1y_{k-1}+a_2y_k=b_k,$$

$$Y_{k+2}+a_2Y_{k+1}+a_2Y_k=b_k.$$

两式相减得 $$d_{k+2}+a_1d_{k+1}+a_2d_k=0,$$

这里 $d_k=y_k-Y_k$.这使得 d_k 是齐次方程的解,所以 $d_k=c_1u_k+c_2v_k$.最后得到 $y_k=c_1u_k+c_2v_k+Y_k$,这就是所需的结论.

18.28 利用前一题,要求非齐次方程的一切解,我们可以求出一个这种特解再加上相应齐次方程的解.用这种方法来解 $y_{k+2}-y_{k+1}-y_k=Ax^k$.

解 ☞ 当右端项 b_k 是幂函数时,通常能找到一个本身也是幂函数的解.

我们在这里试图决定常数 C,使得 $Y_k=Cx^k$.

把 Y_k 代入方程得到 $Cx^k(x^2-x-1)=Ax^k$,因此 $C=A/(x^2-x-1)$.于是它的解可表为

$$Y_k=c_1\left(\frac{1+\sqrt{5}}{2}\right)^k+c_2\left(\frac{1-\sqrt{5}}{2}\right)^k+\frac{Ax^k}{x^2-x-1}.$$

如果 $x^2-x-1=0$,此法无效.

18.29 当 $x^2-x-1=0$ 时,上述问题如何能求出特解 Y_k.

解 ☞ 我们试图确定 C,使得 $Y_k=Ckx^k$.

代入方程得到 $Cx^k[(k+2)x^2-(k+1)x-k]=Ax^k$.由此可得 $C=A/(2x^2-x)$,于是 $Y_k=Akx^k/(2x^2-x)$.

18.30 对哪一种右端项 b_k,可求得基本解 Y_k?

解 ☞ 每当 b_k 是幂函数或正弦函数和余弦函数时,解 Y_k 就有类似的性质.表18.1对此说明得更清楚.如果表 18.1 中给的 Y_k 包含相应齐次方程的解,那么用 k 乘以这个 Y_k 直到不再包含这种解.我们将进一步给出例子来说明这种方法的有效性.

表 18.1

b_k	Y_k
Ax^k	Cx^k
k^n	$C_0+C_1k+C_2k^2+\cdots+C_nk^n$
$\sin Ak$ 或 $\cos Ax$	$C_1\sin Ak+C_2\cos Ak$
k^nx^k	$x^k(C_0+C_1k+C_2k^2+\cdots+C_nk^n)$
$x^k\sin Ak$ 或 $x^k\cos Ak$	$x^k(C_1\sin Ak+C_2\cos Ak)$

补 充 题

18.31 给定 $y_{k+1}=ry_k+k$ 及 $y_0=A$,直接计算 $y_1,\cdots,y_4$.再找出解函数的性质.

18.32 给定 $y_{k+1}=-y_k+4$ 及 $y_0=1$,直接计算 $y_1,\cdots,y_4$.解函数的性质是什么? 对任意的 y_0,你能否找到解的性质.

18.33 如果债务以正常的付款额分期付款,利率为 i,未还结算是 P_k,这里 $P_{k+1}=(1+i)P_k-R$.初始债务是 $P_0=A$,证明 $P_k=A(1+i)^k-R\dfrac{(1+i)^k-1}{i}$.并证明为了刚好在 n 次还款后使P_k 等于零($P_n=0$),我们必须取 $R=Ai/[1-(1+i)^{-n}]$.

18.34 证明带有初始条件 $y_0=2$ 的差分方程 $y_{k+1}=(k+1)y_k+(k+1)!$ 有解 $y_k=k!\ (k+2)$.

18.35 求解 $y_{k+1}=ky_k+2^kk!$, $y_0=0$.

18.36 用题 18.5 中的 Horner 方法计算 $p(x)=1+x+x^2+\cdots x^6$ 在 $x=\dfrac{1}{2}$ 处的值.

18.37 把 Horner 方法用于 $p(x)=x-x^3/3!+x^5/5!-x^7/7!+x^9/9!$.

18.38 对 $k>0$,证明方程$(k+1)y_{k+1}+ky_k=2k-3$ 有解 $y_k=1-2/k$.

18.39 证明非线性方程 $y_{k+1}=y_k/(1+y_k)$有解 $y_k=C/(1+Ck)$.

18.40 求解方程 $\Delta y_k=(1/k-1)y_k$.初始条件是 $y_1=1$.

18.41 根据题 18.11 中的结果计算 $\psi^{(3)}(0)$, $\psi^{(3)}(1)$和 $\psi^{(3)}(2)$,并对整数自变量给出一般结果.

18.42 借助于 ψ 函数计算 $\sum\limits_{k=1}^{\infty}1/k(k+2)$.

18.43 利用题 18.41 计算 $\sum\limits_{k=1}^{\infty}1/k^2(k+2)^2$.

18.44 根据级数定义,用加速方法计算 $\psi\left(\dfrac{1}{2}\right)$到三位小数.然后根据 $\Delta\psi=1/(x+1)$来计算 $\psi\left(\dfrac{3}{2}\right)$和 $\psi\left(\dfrac{-1}{2}\right)$.

18.45 根据上题,当 x 趋于 -1 时,$\psi(x)$的性态如何?

18.46 计算 $\sum\limits_{x=1}^{\infty}1/P_3(x)$.这里 $P_3(x)$是三次 Legendre 多项式.

18.47 计算 $\sum\limits_{x=1}^{\infty}1/T_3(x)$.这里 $T_3(x)$是三次 chebyshev 多项式.

18.48　计算 $\sum_{x=1}^{\infty} 1/P_4(x)$. 这里 $P_4(x)$是四次 Legendre 多项式.

18.49　给定方程 $y_{k+2}+3y_{k+1}+2y_k=0$ 和初始条件 $y_0=2, y_1=1$, 直接计算 $y_2, \cdots y_{10}$.

18.50　用题 18.16 中的方法来解前题.

18.51　证明方程 $y_{k+2}-4y_{k+1}+4y_k=0$ 的解是 $y_k=2^k(c_1+c_2k)$, 这里 c_1, c_2 是任意常数.

18.52　求方程 $y_{k+2}-y_k=0$ 的解族. 再求出满足初始条件 $y_0=0, y_1=1$ 的解.

18.53　求方程 $y_{k+2}-7y_k+12y_k=\cos k$ 的解满足 $y_0=0, y_1=0$.

18.54　求方程 $4y_{k+2}+4y_{k+1}+y_k=k^2$ 的解满足 $y_0=0, y_1=0$.

18.55　证明方程 $y_{k+2}-2y_{k+1}+2y_k=0$ 的解是

$$y_k = c_1(\sqrt{2})^k\sin\frac{\pi k}{4} + c_2(\sqrt{2})^k\cos\frac{\pi k}{4}.$$

18.56　求方程 $2y_{k+2}-5y_{k+1}+2y_k=0$ 的解满足初始条件 $y_0=0, y_1=1$.

18.57　求方程 $y_{k+2}+6y_{k+1}+25y_k=2^k$ 的解满足 $y_0=0, y_1=0$.

18.58　求方程 $y_{k+2}-4y_{k+1}+4y_k=\sin k+2^k$ 的解满足 $y_0=y_1=0$.

18.59　a 取什么值时方程 $y_{k+2}-2y_{k+1}+(1-a)y_k=0$ 的解呈振荡性态.

18.60　求方程 $y_{k+2}-2y_{k+1}-3y_k=P_2(k)$的解满足 $y_0=y_1=0$. 这里 $P_2(k)$是二次 Legendre 多项式.

18.61　当 $0<a<1$ 或 $a=1$ 或 $a>1$ 时, 方程 $y_{k+2}-2ay_{k+1}+ay_k=0$ 的解的性态如何?

18.62　证明通过变量代换 $Q_k=y_{k+1}/y_k$ 能把非线性方程 $Q_{k+1}=a-b/Q_k$ 变换为线性方程 $y_{k+2}-ay_{k+1}-by_k=0$.

18.63　证明对于偶数 N, 方程 $y_{k+2}-y_k=0$ 不存在满足边界条件 $y_0=0, y_N=1$ 的解.

18.64　证明前题的方程有无穷多个解满足是 $y_0=y_N=0$.

18.65　证明对于奇数 N, 方程 $y_{k+2}-y_k=0$ 恰有一个解满足 $y_0=0, y_N=1$. 求出这个解. 并证明仅有一个解满足 $y_0=y_N=0$, 即 $y_k\equiv 0$.

第十九章　微分方程

经典问题

解微分方程是数值分析的主要问题之一.这是因为各种各样的应用导出的微分方程很少能用解析方法求解.经典的初值问题是求一函数 $y(x)$满足一阶微分方程 $y'=f(x,y)$,并取初值 $y(x_0)=y_0$.已设计了各种各样的方法来求这问题的近似解.其中许多方法也被用来处理高阶问题.本章集中讨论这个问题的解法.

1. 首先提出**等倾线法**.根据 $y'(x)$是解曲线的斜率的几何解释,就给出了全解族的定性图.函数 $f(x,y)$确定了每点的斜率.这个所谓的"方向场"决定了解曲线的性质.
2. **历史上的 Euler 方法**是对变量 x_k 应用差分方程
$$y_{k+1}=y_k+hf(x_k,y_k)$$
计算一组离散集合上 y_k 的值,这里 $h=x_{k+1}-x_k$.这是一种明显的不太精确的关于 $y'=f(x,y)$的近似.
3. **计算解的更有效的方法**于是就发展起来了.多项式近似是许多众所周知的算法的基础.除了某些级数方法,多项式近似方法实际计算的是对应于一组离散变量 x_k 的y_k 值的序列.和 Euler 方法一样,大多数方法等价于用差分方程代替所给的微分方程.所得到的特殊的差分方程取决于近似多项式的选取.
4. 大量地应用 **Taylor 级数**.假设 $f(x,y)$是解析函数可以求出 $y(x)$的各阶导数,并可通过标准的 Taylor 公式写出关于 $y(x)$的级数.有时一个级数可以对付所有感兴趣的变量.在其他问题中,一个级数可能收敛太慢而不能对所有感兴趣的变量产生所需要的精度,这时可用在不同点上展开的 Taylor 级数.任何这种级数的最终截断误差表明解将被 Taylor 多项式所近似.
5. 龙格-库塔(Runge-Kutta)方法是为了避免 Taylor 方法中可能包含的关于高阶导数的计算而发展起来的.用给定函数 $f(x,y)$的其他值来代替这些导数.这种方法可产生 Taylor 多项式的两倍精度.最常见的公式是
$$k_1=hf(x,y),$$
$$k_2=hf\left(x+\frac{1}{2}h,y+\frac{1}{2}k_1\right),$$
$$k_3=hf\left(x+\frac{1}{2}h,y+\frac{1}{2}k_2\right),$$
$$k_4=hf(x+h,y+k_3),$$
$$y(x+h)\approx y(x)+\frac{1}{6}(k_1+2k_2+2k_3+k_4).$$
但是有许多变形.
6. **预测-校正方法**包括先用一个公式来预测下一个 y_k 值,然后再用一个更精确的校正公式,于是就可以不断改进精度.虽然稍为复杂,但是这种方法具有通过对每一个 y_k 值的不断逼近就能作出误差估计的优点.简单预测-校正步是
$$y_{k+1}\approx y_k+hy'_k,$$
$$y_{k+1}\approx y_k+\frac{1}{2}h(y'_k+y'_{k+1}).$$
预测步就是 Euler 公式,而校正步称为修正 Euler 公式.由 $y'_k=f(x_k,y_k)$和 $y'_{k+1}=f(x_{k+1},y_{k+1})$,所以预测步先估计 y_{k+1},然后由这个估计求出y'_{k+1}的值,再去修

正 y_{k+1}. 可以继续不断地得到y'_{k+1}和 y_{k+1}的校正，直到取得满意的结果.

7. **Milne 方法**采用预测-校正步

$$y_{k+1} \approx y_{k-3} + \frac{4h}{3}(2y'_{k-2} - y'_{k-1} + 2y'_k),$$

$$y_{k+1} \approx y_{k-1} + \frac{h}{3}(y'_{k+1} + 4y'_k + y'_{k-1}).$$

容易看出其中用到 Simpson 法则，它预先需要四个先行的值($y_k, y_{k-1}, y_{k-2}, y_{k-3}$). 这些值须由另外的方法获得.(通常是 Taylor 级数)

8. **Adams 方法**采用预测-校正步

$$y_{k+1} \approx y_k + \frac{h}{24}(55y'_k - 59y'_{k-1} + 37y'_{k-2} - 9y'_{k-3})$$

$$y_{k+1} \approx y_k + \frac{h}{24}(9y'_{k+1} + 19y'_k - 5y'_{k-1} + y'_{k-2})$$

和 Milne 方法一样，它需要四个先行的值.

误差

截断误差 当用部分和来近似无穷级数的值时就产生了截断误差. 这或许是这一术语原本的用法. 现在它已被用得更自由了. 当差分方程代替微分方程时，随着每向前一步从 k 到 $k+1$ 就产生了局部截断误差，然后这些局部截断误差用某种不清楚的方式产生累积误差或全局截断误差. 不太可能通过微分方程算法的实现来探索误差的增长，但可以作某种粗略的估计.

收敛方法 当一种方法不断改进时(级数的项用得越来越多，或者相邻两变量的间隔越来越小)就产生收敛于准确解的近似解序列，这种方法就是收敛性方法. 可以证明在适当条件下，Taylor, Runge-Kutta 和某些预测-校正方法是收敛的. 收敛性证明只处理截断误差，不顾及问题的舍入误差.

舍入误差 不用多说所有这些方法都出现舍入误差，有时还很重要. 它比截断误差更难以捉摸，而且仅得到很有限的分析.

相对误差 近似的相对误差就是误差与准确解之比，一般比误差本身有更大的兴趣. 因为如果解变得较大，那么就可能容许较大的误差. 更为重要的是，如果准确解减小，那么误差必须同样减小，否则它将淹没解，而将使得计算结果无意义. 简单问题 $y' = Ay$, $y(0) = A$，其准确解是 $y = e^{Ax}$，常用作我们各种方法中探索相对误差性质的试验. 我们希望这种方式得到的信息和把同样的方法用于一般方程 $y' = f(x, y)$有所联系. 这似乎是乐观的，但是误差的研究有其局限性.

稳定方法 是一种希望通过初值使相对误差保持有界的方法. 这是一个很强的要求而且很难证明. 还有，一种方法可能对有些方程是稳定的，而对另一些方程是不稳定的. 仅能提供部分结论，特别是对于方程 $y' = Ay$.

误差控制 涉及到一步一步地测量局部截断误差，并用此信息来确定当前的步长是否合适. 对于预测-校正方法，用预测和校正值就能作出实际误差估计. 在 Runge-Kutta 方法中，用双步长的平行计算得出的误差估计很像在自适应积分中. 和那里一样，这里的目的是用最小的努力达到指定精度的最终结果.

题　解

等斜线方法

19.1 用等斜线方法来确定 $y'(x) = xy^{1/3}$的解的性态.

解 ☞ 当然可以用初等方法来解这个方程. 但是我们把它作为对各种近似方法的试验. 等斜线方

法是根据曲线族 $y'(x)$ = 常数本身不是解，但却有助于确定解的特性. 这个例子中，等斜线是曲线族 $xy^{1/3} = M$，这里 M 是 $y'(x)$ 的常数值. 这些曲线中的一部分描绘(虚线的)在图 19.1 中，M 值已表出. 在微分方程的解穿过这些等斜线中的一条处，它一定以那条等斜线的 M 为其斜率. 图 19.1 中也包含几条解曲线(实线). 其余的至少能大致描出.

等斜线方法的目标不是精度而是解族的一般性质. 例如都有关于两轴的对称性. 经过(0,0)的解及其上面的解都是 U 形. 在它下面的解更是异于寻常. 沿 $y=0$ 不同的解能在一起，甚至一个解能包含一段 x 轴. 这样的解可能沿着下降弧进入(0,0)，再沿着 x 轴到(2,0)，然后再向上. 如图 19.2 所示. 可能有数不尽的直线和弧结合在一起. 当力图计算精确解时，这类信息常是有用的指导.

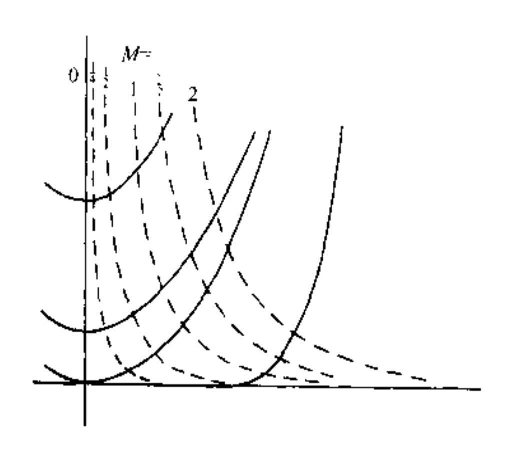


图 19.1

图 19.2

Euler 方法

19.2 叙述用最简单的 Euler 方法计算

$$y' = f(x,y) = xy^{1/3} \qquad y(1) = 1$$

的解.

解 这或许是把等斜线转换成一个计算格式的最初的方法，它用公式

$$y_{k+1} - y_k = \int_{x_k}^{x_{k+1}} y' \mathrm{d}x \approx hy'_k,$$

这相当于认为在 x_k 和 x_{k+1} 之间 y' 是常数. 它也相当于 Taylor 级数的线性部分. 因此如果 y_k 和 y'_k 精确地知道，那么在 y_{k+1} 的误差将是 $\frac{1}{2}h^2y^{(2)}(\xi)$. 它称为局部截断误差，这是因为它在 x_k 到 x_{k+1} 这一步中产生的，由于它很大，为了高精度就需要相当小的增量 h.

这个公式很少用于实际，但用来表明今后任务的性质和某些必须面对的困难. 对于 x_0, $y_0=1$, 取 $h=0.01$ 三次使用这一 Euler 公式得到

$$y_1 \approx 1 + (0.01)(1) = 1.0100$$

$$y_2 \approx 1.0100 + (0.01)(1.01)(1.0033) \approx 1.0201$$

$$y_3 \approx 1.0201 + (0.01)(1.02)(1.0067) \approx 1.0304$$

在 $x=1$ 附近我们有 $y^{(2)} = y^{1/3} + \frac{1}{3}xy^{-2/3}\cdot(xy^{1/3}) \approx \frac{4}{3}$，在每一步就产生大约 0.00007 的截断误差. 经过三次这种误差，第四位小数已值得怀疑. 如果我们希望较高的精度，增量 h 就必须比较小. 在图 19.3 中进一步说明截断误差的累积. 在那里由计算得到的点已参与测出解曲线. 我们的近似相当于对方程的各个解沿着切线方向. 近似解最终趋于沿着解曲线的凸出的一边. 还要注意 Euler 公式就是一阶非线

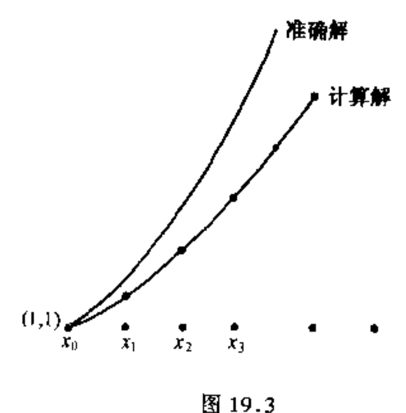


图 19.3

性差分方程：$y_{k+1}=y_k+hx_ky_k^{1/3}$.

19.3 取 $h=0.10, 0.05$ 和 0.01，用 Euler 方法所得结果与精确解 $y=[(x^2+2)/3]^{3/2}$ 进行比较，说明收敛性概念.

解 ☞ 收敛是指当区间长 h 趋于零时，近似解的改进，一个不收敛的方法作为一种差分格式是值得怀疑的. 后面将证明所引进各种格式的收敛性. 但作为旁证，由 Euler 方法得到的表 19.1 中的数据可供参考. 只包括对于整数变量 x 的值. 为简单起见，其余的都被删掉了.

注意通过每一行，确实都趋于精确值. 用较小的区间意味着更多的计算. 例如在末行，值 25.96 是经过 50 步得到的，而值 26.89 需要 500 步，额外的工作带来了改进，这才是公平的. 当 h 趋于零时，计算将更长. 我们希望其结果作为极限趋于准确值.

表 19.1

x	$h=0.10$	$h=0.05$	$h=0.01$	准确
1	1.00	1.00	1.00	1.00
2	2.72	2.78	2.82	2.83
3	6.71	6.87	6.99	7.02
4	14.08	14.39	14.63	14.70
5	25.96	26.48	26.89	27.00

这就是收敛性概念，不用说舍入误差将限止所获精度，但它们不在收敛性问题内.

Taylor 方法

19.4 用局部 Taylor 级数方法，对自变量直到 $x=5$ 求 $y'=xy^{1/3}$, $y(1)=1$ 的解，精确到三位数.

解 ☞ 一般讲这个方法包括用 $p(x+h)$ 代替 $y(x+h)$，这里的 $p(x)$ 是关于变量 x 的 Taylor 多项式. 我们可以直接写出

$$y(x+h)\approx y(x)+hy'(x)+\frac{1}{2}h^2y^{(2)}(x)+\frac{1}{6}h^3y^{(3)}(x)+\frac{1}{24}h^4y^{(4)}(x),$$

它的局部截断误差 $E=h^5y^{(5)}(\xi)/120$.

由微分方程可算出高阶导数

$$y^{(2)}(x)=\frac{1}{3}x^2y^{-1/3}+y^{1/3},$$

$$y^{(3)}(x)=-\frac{1}{9}x^3y^{-1}+xy^{-1/3},$$

$$y^{(4)}(x)=\frac{1}{9}x^4y^{-5/3}-\frac{2}{3}x^2y^{-1}+y^{-1/3},$$

初始条件 $y(1)=1$ 已给定. 当 $x=1$ 和 $h=0.1$ 时，就得到

$$y(1+0.1)\approx 1+0.1+\frac{2}{3}(0.1)^2+\frac{4}{27}(0.1)^3+\frac{1}{54}(0.1)^4\approx 1.10682.$$

下一步在 $x=1.1$ 处用 Taylor 公式得到

$$y(1.1+0.1)\approx 1.22788\qquad y(1.1-0.1)\approx 1.00000$$

其中第二个用作精度检查，因它再现了我们第一个结果达到五位精度.（这与第十四章中对误差函数积分所用的步骤是相同的.）继续这种方法就得到了表 19.2 所示结果. 再次用准确解作比较. 虽然用的是 $h=0.1$，仅列出 $x=1(0.5)5$ 的值. 注意误差比 Euler 方法取 $h=0.01$ 所产生的误差要小得多. Taylor 方法是一种收敛比较迅速的算法.

表 19.2

x	Taylor 结果	准确结果	误差
1.0	1.00000	1.00000	
1.5	1.68618	1.68617	-1

续表

x	Taylor 结果	准确结果	误差
2.0	2.82846	2.82843	-3
2.5	4.56042	4.56036	-6
3.0	7.02123	7.02113	-10
3.5	10.35252	10.35238	-14
4.0	14.69710	14.69694	-16
4.5	20.19842	20.19822	-20
5.0	27.00022	27.00000	-22

19.5 用 Taylor 方法求 $y' = -xy^2, y(0)=2$ 的解.

解 ☞ 能用上述题的方法，但将说明其变化，本质上是一种待定系数法. 一开始就假定收敛，我们写出 Taylor 级数 $y(x) = \sum_{i=0}^{\infty} a_i x^i$ 于是

$$y^2(x) = \left(\sum_{i=0}^{\infty} a_i x^i\right)\left(\sum_{j=0}^{\infty} a_j x^j\right) = \sum_{k=0}^{\infty}\left(\sum_{i=0}^{k} a_i a_{k-i}\right)x^k, \qquad y'(x) = \sum_{i=0}^{\infty} i a_i x^{i-1}$$

代入微分方程并对和数的下标稍作改变得到

$$\sum_{j=0}^{\infty}(j+1)a_{j+1}x^j = -\sum_{j=1}^{\infty}\left(\sum_{i=0}^{j-1} a_i a_{j-1-i}\right)x^j.$$

比较 x^j 的系数得到 $a_1 = 0$ 和

$$(j+1)a_{j+1} = -\sum_{i=0}^{j-1} a_i a_{j-1-i} \qquad j = 1,2,\cdots.$$

初始条件使得 $a_0 = 2$，然而递推地得到

$$a_2 = -\frac{1}{2}a_0^2 = -2, \qquad a_6 = -\frac{1}{6}(2a_0a_4 + 2a_1a_3 + a_2^2) = -2,$$

$$a_3 = -\frac{1}{3}(2a_0a_1) = 0, \qquad a_7 = -\frac{1}{7}(2a_0a_5 + 2a_1a_4 - 2a_2a_3) = 0,$$

$$a_4 = -\frac{1}{4}(2a_0a_2 - a_1^2) = 2, \qquad a_8 = -\frac{1}{8}(2a_0a_6 + 2a_1a_5 + 2a_2a_4 + a_3^2) = 2,$$

$$a_5 = -\frac{1}{5}(2a_0a_3 - 2a_1a_2) = 0,$$

等等. 递推可程序化，所以系数可根据需要而自动计算. 所标明的级数是

$$y(x) = 2(1 - x^2 + x^4 - x^6 + x^8 \quad \cdots).$$

因为容易求得精确解是 $y(x) = 2/(1+x^2)$，所以得到上述级数就不足为怪了. 这个方法经常有用，其基本假设是解确实具有级数表达式，在这种情形下，仅当 $-1<x<1$ 时级数收敛. 当 $-\frac{1}{2}<x<\frac{1}{2}$ 时只需要六项给出三位精度. 在前题中为了计算每一个值用了新的 Taylor 多项式. 在这里一个这样的多项式就足够了. 问题是范围和需要的精度. 例如，直到 $x=5$，前一方法都能使用. 进一步对照我们还注意到在题 19.4 中用到固定项数的多项式，而且没有明显地发生收敛问题. 这里假设 $y(x)$ 在感兴趣的区间内是解析的，在题 19.5 中我们把整个级数引进微分方程.

Runge-Kutta 方法

19.6 确定系数 a, b, c, d, m, n 和 p 使 Runge-Kutta 公式

$$k_1 = hf(x, y),$$

$$k_2 = hf(x + mh, y + mk_1),$$

$$k_3 = hf(x + nh, y + nk_2),$$

$$k_4 = hf(x + ph, y + pk_3),$$

$$y(x+h) - y(x) \approx ak_1 + bk_2 + ck_3 + dk_4$$

和直到 h^4 的 Taylor 级数一样. 注意最后一个式子尽管不是多项式近似但接近四次 Taylor 多项式.

解 ☞ 我们从把 Taylor 级数表成便于比较的形式开始. 令

$$F_1 = f_x + ff_y, \qquad F_2 = f_{xx} + 2ff_{xy} + f^2 f_{yy},$$
$$F_3 = f_{xxx} + 3ff_{xxy} + 3f^2 f_{xyy} + f^3 f_{yyy},$$

然后对方程 $y' = f(x, y)$ 两边求导得

$$y^{(2)} = f_x + f_y y' = f_x + f_y f = F_1,$$
$$y^{(3)} = f_{xx} + 2ff_{xy} + f^2 f_{yy} + f_y(f_x + ff_y) = F_2 + f_y F_1,$$
$$\begin{aligned} y^{(4)} &= f_{xxx} + 3ff_{xxy} + 3f^2 f_{xyy} + f^3 f_{yyy} + f_y(f_{xx} + 2ff_{xy} + f^2 f_{yy}) \\ &\quad + 3(f_x + ff_y)(f_{xy} + ff_{yy}) + f_y^2(f_x + ff_y) \\ &= F_3 + f_y F_2 + 3F_1(f_{xy} + ff_{yy}) + f_y^2 F_1. \end{aligned}$$

这使得 Taylor 级数可写成

$$\begin{aligned} y(x+h) - y(x) &= hf + \frac{1}{2}h^2 F_1 + \frac{1}{6}h^3(F_2 + f_y F_1) \\ &\quad + \frac{1}{24}h^4[F_3 + f_y F_2 - 3(f_{xy} + ff_{yy})F_1 + f_y^2 F_1] + \cdots. \end{aligned}$$

现转向各个 k 值,类似的计算得到

$$k_1 = hf,$$
$$k_2 = h\left[f + mhF_1 + \frac{1}{2}m^2h^2F_2 + \frac{1}{6}m^3h^3F_3 + \cdots\right],$$
$$\begin{aligned} k_3 = h\Big[&f + nhF_1 + \frac{1}{2}h^2(n^2F_2 + 2mnf_yF_1) \\ &+ \frac{1}{6}h^3(n^3F_3 + 3m^2nf_yF_2 + 6mn^2(f_{xy} + ff_{yy})F_1) + \cdots\Big], \end{aligned}$$
$$\begin{aligned} k_4 = h\Big[&f + phF_1 + \frac{1}{2}h^2(p^2F_2 + 2npf_yF_1), \\ &+ \frac{1}{6}h^3(p^3F_3 + 3n^2pf_yF_2 + 6np^2(f_{xy} + ff_{yy})F_1 + 6mnpf_y^2F_1)\cdots\Big]. \end{aligned}$$

把这些一起代入 Runge-Kutta 公式的最后一式并重新组合得到

$$\begin{aligned} y(x+h) - y(x) &= (a + b + c + d)hf + (bm + cn + dp)h^2F_1 \\ &\quad + \frac{1}{2}(bm^2 + cn^2 + dp^2)h^3F_2 + \frac{1}{6}(bm^3 + cn^3 + dp^3)h^4F_3 \\ &\quad + (cmn + dnp)h^3f_yF_1 + \frac{1}{2}(cm^2n + dn^2p)h^4f_yF_2 \\ &\quad + (cmn^2 + dnp^2)h^4(f_{xy} + ff_{yy})F_1 + dmnph^4f_y^2F_1 + \cdots. \end{aligned}$$

与 Taylor 级数比较后得到八个条件

$$a + b + c + d = 1, \qquad cmn + dnp = \frac{1}{6},$$
$$bm + cn + dp = \frac{1}{2}, \qquad cmn^2 + dnp^2 = \frac{1}{8},$$
$$bm^2 + cn^2 + dp^2 = \frac{1}{3}, \qquad cm^2n + dn^2p = \frac{1}{12},$$
$$bm^3 + cn^3 + dp^3 = \frac{1}{4}, \qquad dmnp = \frac{1}{24}.$$

这八个方程含七个未知量,实际上有一个是多余的,其典型解集是

$$m = n = \frac{1}{2}, \qquad p = 1, \qquad a = d = \frac{1}{6}, \qquad b = c = \frac{1}{3},$$

这样就导出了 Runge-Kutta 公式

$$k_1 = hf(x, y), \qquad k_2 = hf\left(x + \frac{1}{2}h, y + \frac{1}{2}k_1\right),$$
$$k_3 = hf\left(x + \frac{1}{2}h, y + \frac{1}{2}k_2\right), \qquad k_4 = hf(x + h, y + k_3),$$
$$y(x+h) \approx y(x) + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

颇有兴趣的是注意到与 y 无关的 $f(x, y)$ 这就导致用于 $y'(x) = f(x)$ 的 Simpson 法则.

19.7　和 Taylor 方法相比,Runge-Kutta 公式的优点是什么?

解 ☞ 尽管和四次 Taylor 多项式的近似程度相同，这些公式并不需要像 Taylor 方法那样，预先计算 $y(x)$ 的较高阶导数. 由于出现在应用中的微分方程常常是复杂的，所以计算导数可能很麻烦. Runge-Kutta 方法用计算不同位置的 $f(x,y)$ 来代替，而这一函数出现在所给方程中. 这个方法应用很广泛.*

19.8 对 $y'=f(x,y)=xy^{1/3}$, $y(1)=1$ 应用 Runge-Kutta 公式.

解 ☞ 取 $x_0=1$ $h=0.1$，我们得到

$$k_1=(0.1)f(1,1)=0.1, \qquad k_3=(0.1)f(1.05,1.05336)\approx 0.10684,$$

$$k_2=(0.1)f(1.05,1.05)\approx 0.10672, \qquad k_4=(0.1)f(1.1,1.10684)\approx 0.11378.$$

由此我们算得

$$y_1=1+\frac{1}{6}(0.1+0.21344+0.21368+0.11378)\approx 1.10682.$$

这一步完成，我们用 x_1 和 y_1 代替 x_0 和 y_0 开始另一步，并以这种方式继续下去. 因为这种方法与直到 h^4 的 Taylor 级数一样，很自然要求所得结果类似于 Taylor 方法所得的结果. 表 19.3 作了一些比较，我们在最后两位发现了差别，这就通过事实部分地说明了两种方法的局部截断误差不完全相同. 两者都具有 Ch^5 的形式，但是因子 C 不同. 还有即使在代数上完全相同的不同算法中，其舍入误差通常是不同的. 当然这两种方法是不相同的. Runge-Kutta 方法的优点是明显的.

表 19.3

x	Taylor	Runge-Kutta	准确
1	1.00000	1.00000	1.00000
2	2.82846	2.82843	2.82843
3	7.02123	7.02113	7.02113
4	14.69710	14.69693	14.69694
5	27.00022	26.99998	27.00000

19.9 叙述 Runge-Kutta 公式的各种变形.

解 ☞ 不难证明

$$y(x+h)=y(x)+hf\left(x+\frac{1}{2}h,y+\frac{1}{2}hf(x,h)\right),$$

其中 y 表示 $y(x)$，与到二次项的 Taylor 级数相同. 因此称为二阶 Runge-Kutta 方法. 同样地

$$k_1=hf(x,y),$$

$$k_2=hf\left(x+\frac{1}{2}h,y+\frac{1}{2}k_1\right),$$

$$k_3=hf(x+h,y-k_1+2k_2),$$

$$y(x+h)=y(x)+\frac{1}{6}(k_1+4k_2+k_3)$$

是三阶的. 还存在其他二阶和三阶方法. 以下一组

$$k_1=hf(x,y),$$

$$k_2=hf\left(x+\frac{1}{2}h,y+\frac{1}{2}k_1\right),$$

$$k_3=hf\left(x+\frac{1}{2}h,y+\frac{1}{4}k_1+\frac{1}{4}k_2\right),$$

$$k_4=hf(x+h,y-k_2+2k_3),$$

$$y(x+h)=y(x)+\frac{1}{6}(k_1+4k_3+k_4)$$

是一种四阶方法，而更奇特地

* 译注：这种方法通常称为间接使用 Taylor 展式法.

$$k_1 = f(x, y),$$
$$k_2 = hf\left(x + \frac{1}{2}h, y + \frac{1}{2}k_1\right),$$
$$k_3 = hf\left(x + \frac{1}{2}h, y + \frac{1}{4}k_1 + \frac{1}{4}k_2\right),$$
$$k_4 = hf(x + h, y - k_2 + 2k_3),$$
$$k_5 = hf\left(x + \frac{2}{3}h, y + \frac{7}{27}k_1 + \frac{10}{27}k_2 + \frac{1}{27}k_4\right),$$
$$k_6 = hf\left(x + \frac{1}{5}h, y + \frac{28}{625}k_1 - \frac{1}{5}k_2 + \frac{546}{625}k_3 + \frac{54}{625}k_4 - \frac{378}{625}k_5\right),$$
$$y(x + h) = y(x) + \frac{1}{24}k_1 + \frac{5}{48}k_4 + \frac{27}{56}k_5 + \frac{125}{336}k_6$$

是五阶的.阶数越高,可能的方法其差异就越大.截断误差就越小.n 阶方法与直到 n 次项的 Taylor 级数一样,因此有截断误差

$$T = cy^{(n+1)}h^{n+1}.$$

这就意味着对于光滑函数,能用较大的 h 进行计算而且很快.建立高阶方法涉及艰苦的代数运算.只有在计算机帮助下才可能进行这种运算.

Taylor 方法的收敛性

19.10　方程 $y' = y$ 和 $y(0) = 1$ 有精确解 $y(x) = e^x$.证明当 p 固定,h 趋于零时,由 Taylor 方法得到的近似值 y_k 收敛于这个精确解.(较熟悉的收敛概念是保持 h 固定而让 p 趋于无穷.)

解　☞ Taylor 方法包括对每一个精确值用

$$Y_{k+1} = Y_k + hY_k' + \frac{1}{2}h^2Y_k^{(2)} + \cdots + \frac{1}{p!}h^pY_k^{(p)}$$

来近似.

对目前的问题所有的导数都相同,就得到

$$Y_{k+1} = \left(1 + h + \frac{1}{2}h^2 + \cdots + \frac{1}{p!}h^p\right)Y_k = rY_k.$$

当 $p = 1$ 时它就退化为 Euler 方法,在任何情形它都是一阶差分方程.它的满足 $Y_0 = 1$ 的解是

$$Y_k = r^k = \left(1 + h + \frac{1}{2}h^2 + \cdots + \frac{1}{p!}h^p\right)^k.$$

但是根据 Taylor 多项式,公式

$$e^h = 1 + h + \frac{1}{2}h^2 + \cdots + \frac{1}{p!}h^p + \frac{h^{p+1}}{(p+1)!}e^{\xi h},$$

ξ 在 0 和 1 之间.现在想到恒等式

$$a^k - r^k = (a - r)(a^{k-1} + a^{k-2}r + \cdots + ar^{k-2} + r^{k-1}),$$

对于 $a > r > 0$ 有

$$a^k - r^k < (a - r)ka^{k-1}.$$

取 $a = e^h$,r 如前,最后的不等式成为

$$0 < e^{kh} - Y_k < \frac{h^{p+1}}{(p+1)!}e^{\xi h}ke^{(k-1)h} < \frac{kh^{p+1}}{(p+1)!}e^{kh}.$$

最后一步是因为 $0 < \xi < 1$.收敛性问题关系到已算得的值对于固定的变量 x,当 h 趋于零时的性态,* 因此我们令 $x_k = kh$,把上面最后的结果写成

$$0 < e^{x_k} - Y_k < \frac{h^p}{(p-1)!}x_ke^{x_k}.$$

---

* 译注:这里特别需要加以补充说明的是,常微分方程数值解在某点 x_k 处的收敛,指的是点收敛,所讨论的点 $x = x_k = kh$ 固定不变.所以当 $h \to 0$ 时就意味着 $k \to \infty$ 必须同时进行,所以收敛的严格定义是 $\lim\limits_{\substack{x = x_k = kh \\ h \to 0 \\ k \to \infty}} y(x_k) = y(x)$,也就是说,对我们感兴趣的点,如果用越来越小的步长,越来越多的步数去逼近它.最后能达到微分方程在这一点的精确值,就是收敛.

现选取步长 h 的一个序列，使得 x_k 在这个每次计算的有限变量集合中无限循环.(最简单的方法就是不断二等分 h).根据上一不等式，所得到的 Y_k 值的序列在固定的变量 x_k 处以 h^p 阶收敛于精确解 e^{x_k}.当然其实际意义是 h 取得越小，所得到计算结果越来越接近精确解.当然这一问题还没有考虑的舍入误差将限制达到的精度.

19.11 在前一题中建立起来的 Taylor 近似的误差，对固定的步长 h，当 k 增加时表现如何？换言之对越来越大的量进行计算时，上述误差表现如何？

解 ☞ 注意因为 h 固定，所以这不是收敛性问题.这是一个由于到 h^p 项的 Taylor 级数的截断误差在不断计算时，所产生的误差是如何积累的问题.我们从最后一个不等式看到误差包含真解作其因子.实际上它是更重要的相对误差，这是因为它关系到在我们的计算值中有效数字的位数，我们得到

$$\text{相对误差} = \left|\frac{e^{x_k} - Y_k}{e^{x_k}}\right| < \frac{h^p}{(p+1)!}x_k.$$

对固定的 h，它关于 x_k 是线性增长的.

19.12 在 $f(x,y)$ 适当的假设下，证明对带有初始条件 $y(x_0)=y_0$ 的一般的一阶方程 $y'=f(x,y)$ 的 Taylor 方法的收敛性.

证 ☞ 这推广了题 19.10 的结果，仍用大写字母 Y 表示近似解、Taylor 方法得到

$$Y_{k+1} = Y_k + hY_k' + \frac{1}{2}h^2Y_k^{(2)} + \cdots + \frac{1}{p!}h^pY_k^{(p)},$$

这里所有的 $Y_k^{(i)}$ 值由微分方程算得.例如

$$Y_k' = f(x_k, Y_k), \qquad Y_k^{(2)} = f_x(x_k, Y_k) + f_y(x_k, Y_k)f(x_k, Y_k) = f'(x_k, Y_k).$$

为简单起见不写出自变量

$$Y_k^{(3)} = f_{xx} + 2f_{xy}f + f_{yy}f^2 + (f_x + f_yf)f_y = f''(x_k, Y_k).$$

这可理解为 f 及其导数在 x_k, Y_k 处求值，Y_k 表示在自变 x_k 处算得的值.其他的 $Y_k^{(i)}$ 由类似的但含有更多项的式子得到.如果我们用 $y(x)$ 表示微分方程的精确解，对于 $y(x_{k+1})$ Taylor 公式提供了类似的表达式

$$y(x_{k+1}) = y(x_k) + hy'(x_k) + \frac{1}{2}h^2y^{(2)}(x_k) + \cdots$$

$$+ \frac{1}{p!}h^py^{(p)}(x_k) + \frac{h^{p+1}}{(p+1)!}y^{(p+1)}(\xi),$$

只要精确解的确存在这些导数.一般地 ξ 在 x_k 和 x_{k+1} 之间.因为 $y'(x)=f(x,y(x))$，我们有

$$y'(x_k) = f(x_k, y(x_k)),$$

两边求导数得

$$y^{(2)}(x_k) = f_x(x_k, y(x_k)) + f_y(x_k, y(x_k))f(x_k, y(x_k)) = f'(x_k, y(x_k)).$$

同样可得

$$y^{(3)}(x_k) = f''(x_k, y(x_k)).$$

如此等等.再相减得

$$y(x_{k+1}) - Y_{k+1} = y(x_k) - Y_k + h[y'(x_k) - Y_k'] + \frac{1}{2}h^2[y^{(2)}(x_k) - Y_k^{(2)}]$$

$$+ \cdots + \frac{1}{p!}h^p[y^{(p)}(x_k) - Y_k^{(p)}] + \frac{h^{p+1}}{(p+1)!}y^{(p+1)}(\xi).$$

现在注意若 $f(x,y)$ 满足利普希茨(Lipschitz)条件

$$|y'(x_k) - Y_k'| = |f(x_k, y(x_k)) - f(x_k, Y_k)| \leqslant L|y(x_k) - Y_k|,$$

我们将进一步假设 $f(x,y)$ 满足

$$|y^{(i)}(x_k) - Y_k^{(i)}| = |f^{(i-1)}(x_k, y(x_k))$$

$$- f^{(i-1)}(x_k, Y_k)| \leqslant L|y(x_k) - Y_k|.$$

假如 $f(x,y)$ 有直到 $p+1$ 阶的连续导数，这就能证明对于 $i=1,2,\cdots,p$ 是正确的.同样的条件也保证精确解 $y(x)$ 有直到 $p+1$ 阶连续导数，即以上所假定的事实，在这些关于 $f(x,y)$ 的假设下，现在令 $d_k = y(x_k) - Y_k$，我们有

$$|d_{k+1}| \leqslant |d_k|\left(1 + hL + \frac{1}{2}h^2L + \cdots + \frac{1}{p!}h^pL\right) + \frac{h^{p+1}}{(p+1)!}B,$$

这里 B 是 $|y^{p+1}(x)|$ 的界.为简单起见,这可写为

$$|d_{k+1}| \leqslant (1+\alpha)|d_k| + \beta.$$

这里 $$\alpha = L\left(h + \frac{1}{2}h^2 + \cdots + \frac{1}{p!}h^p\right), \quad \beta = \frac{h^{p+1}}{(p+1)!}B.$$

现在我们证明 $$|d_k| \leqslant \beta\frac{e^{k\alpha}-1}{\alpha},$$

α 和 β 是正数.因为精确解和近似解都满足初始条件,所以 $d_0=0$.因此最后一个不等式对 $k=0$ 成立,用归纳法证明它,假设它对非负整数 k 成立,我们有

$$|d_{k+1}| \leqslant (1+\alpha)\beta\frac{e^{k\alpha}-1}{\alpha} + \beta = \frac{(1+\alpha)e^{k\alpha}-1}{\alpha}\beta < \frac{e^{(k+1)\alpha}-1}{\alpha}\beta$$

最后一步是因为 $1+\alpha<e^{\alpha}$.因此归纳法成立,即对一切非负整数 k 不等式都成立.由于 $\alpha = Lh + \varepsilon h < Mh$,这里 ε 随着 h 趋于零.我们能用稍大的 M 代替 L.于是得到

$$|y(x_k) - Y_k| \leqslant \frac{h^pB}{(p+1)!}\cdot\frac{e^{M(x_k-x_0)}-1}{M}.$$

用通常的自变量变换 $x_k = x_0 + kh$,所以如同 h_p 一样收敛.

19.13 当计算继续进行到较大的变量 x_k 时,对固定的 h 的误差,题 19.12 的结果说明什么?

解 ☞ 这个结果适用于证明收敛性.但由于不知道精确解,因此不能立即导出相对误差估计,已研究了进一步误差分析和有限步外推.

19.14 Runge-Kutta 方法是否也收敛?

解 ☞ 由于这些方法与到某个位置(在我们的例子中是直到 h^4 的项)的 Taylor 级数是相同的.所以其收敛性证明与刚才提供的对于 Taylor 方法本身的收敛性证明是相同的.细节较复杂.省略.

预测-校正方法

19.15 导出修正 Euler 公式 $y_{k+1} \approx y_k + \frac{1}{2}h(y'_k + y'_{k+1})$ 并求出其局部截断误差.

解 ☞ 这个公式能通过对 y' 的积分用梯形公式而得到,具体如下:

$$y_{k+1} - y_k = \int_{x_k}^{x_{k+1}} y'\,dx \approx \frac{1}{2}h(y'_k + y'_{k-1}).$$

根据问题 14.66,对 y' 的积分用梯形公式的误差将是 $-h^3y^{(3)}(\xi)/12$.这就是局部截断误差.(想起局部截断误差和从 x_k 到 x_{k+1} 这步的近似所引起的误差有关.实际上我们假设 y_k 及以前的值是准确的.)把目前的结果与简单的 Euler 方法相比,我们当然发现目前的误差事实上比较小.这可以认为这是对用梯形公式而不是其他较原始的积分方法的自然回报.注意到在 x_k 和 x_{x+1} 之间不是把 y' 当成常数而是假设 y' 是线性函数,这是很有兴趣的.我们现在在这个区间中把 y' 当成是线性的.所以认为 $y(x)$ 是二次的.

19.16 对 $y' = xy^{1/3}$, $y(1)=1$ 应用修正的 Euler 公式.

解 ☞ 尽管这种方法很少用于严格的计算,但可用来说明预测-校正法的性质.假设 y_k 和 y_k 已知,以下两个方程

$$y_{k+1} \approx y_k + \frac{1}{2}h(y'_k + y'_{k+1}), \qquad y'_{k+1} = f(x_{k+1}, y_{k+1})$$

可用来确定 y_{k+1} 和 y'_{k+1}.用和第二十五章中提出的求方程根的方法十分相似的迭代算法.从 $k=0$ 开始,逐次迭代.这个算法就产生了序列 y_k 和序列 y'_k.同样有趣的是回想起求解前题中的注释,我们在 x_k 之间的 $y(x)$ 是二次的.因此对 $y(x)$ 的整个近似可认为是一串抛物线段.$y(x)$ 和 $y'(x)$ 都连续,而 $y''(x)$ 在"角点"(x_k, y_k) 处有跳跃.

为引发我们算法的每一向前步,先用 Euler 公式作预测,它提供了 y_{k+1} 的第一个估计,这里取 $x_0=1$, $h=0.05$,它给出

$$y(1.05) \approx 1 + (0.05)(1) = 1.05$$

然后由微分方程提供给我们

$$y'(1.05) \approx (1.05)(1.016) \approx 1.0661$$

现在修正 Euler 公式可以用来作校正,得到

$$y(1.05) \approx 1 + (0.025)(1 + 1.0661) \approx 1.05165$$

微分方程用这新值把 $y'(1.05)$ 修正为 1.0678，其后再应用校正步

得到 $$y(1.05) \approx 1 + (0.025)(1 + 1.0678) \approx 1.0517$$

再循环一次又得到这个四位小数的值. 因此停止. 与微分方程一起校正公式的这种迭代使用是预测-校正方法的核心. 假定方法收敛，我们就迭代到收敛.（证明见题 29.29）. 再从单独使用预测公式开始，进行下一个向前步. 因为现在得到了更有功效的预测-校正公式，所以我们不再进一步继续目前的计算. 然而，注意在最后处得到的结果仅是两个单位太小了. 这就证明了我们的校正公式比简单的 Euler 预测步更精确. 当 $h = 0.01$ 时，后者很少产生四位精度. 现在将建立更有功效的预测-校正组合.

19.17 导出"预测"公式 $y_{k+1} \approx y_{k-3} + \frac{4}{3}h(2y'_{k-2} - y'_{k-1} + 2y'_k)$.

解 ☞ 以前（第十四章）我们曾在整个配置区间积分配置多项式（Cotes 公式），也曾仅在区间的一部分积分（端点校正公式）. 虽然较麻烦些，第二个方法得到更精确的结果. 现在我们在比配置区间更大的区间上积分配置多项式. 不必大惊小怪，由此所得到的公式将多少降低精度，但它将起重要作用. 多项式

$$p_k = y'_0 + k\frac{y'_1 - y'_{-1}}{2} + k^2\frac{y'_1 - 2y'_0 + y'_{-1}}{2}$$

对 $k = -1, 0, 1$ 满足 $p_k = y'_k$. 它是 $y'(x)$ 的以二次 Stirling 公式的形式，即抛物线的配置多项式. 从 $k = -2$ 到 $k = 2$ 积分得

$$\int_{-2}^{2} p_k \mathrm{d}k = 4y'_0 + \frac{8}{3}(y'_1 - 2y'_0 + y'_{-1}) = \frac{4}{3}(2y'_1 - y'_0 + 2y'_{-1}).$$

经过通常的自变量变换 $x = x_0 + kh$，它就变成

$$\int_{x_{-2}}^{x_2} p(x)\mathrm{d}k = \frac{4}{3}h(2y'_1 - y'_0 + 2y'_{-1}).$$

由于我们把 $p(x)$ 当成 $y'(x)$ 的一种近似

$$\int_{x_{-2}}^{x_2} y'(x)\mathrm{d}x = y_2 - y_{-2} \approx \frac{4}{3}h(2y'_1 - y'_0 + 2y'_{-1}),$$

因为在其他区间可进行同样的讨论. 下标都可以增加 $k-1$，就得到了所需要的预测公式. 之所以这样称谓，是因为它允许对较小的变量的数据来预测 y_2.

19.18 这个预测步的局部截断误差是什么?

解 ☞ 它可以用 Taylor 级数方法来估计，取零为临时参照点

$$y_k = y_0 + (kh)y'_0 + \frac{1}{2}(kh)^2y_0^{(2)} + \frac{1}{6}(kh)^3y_0^{(3)} + \frac{1}{24}(kh)^4y_0^{(4)} + \frac{1}{120}(kh)^5y_0^{(5)} + \cdots.$$

从它可得到

$$y_2 - y_{-2} = 4hy'_0 + \frac{8}{3}h^3y_0^{(3)} + \frac{8}{15}h^5y_0^{(5)} + \cdots.$$

微商得到

$$y'_k = y'_0 + (kh)y_0^{(2)} + \frac{1}{2}(kh)^2y_0^{(3)} + \frac{1}{6}(kh)^3y_0^{(4)} + \frac{1}{24}(kh)^4y_0^{(5)} + \cdots.$$

由此可得 $$2y'_1 - y'_0 + 2y'_{-1} = 3y'_0 + 2h^2y_0^{(3)} + \frac{1}{6}h^4y_0^{(5)} + \cdots.$$

因此局部截断误差是

$$(y_2 - y_{-2}) - \frac{4}{3}h(2y'_1 - y'_0 + 2y'_{-1}) = \frac{14}{45}h^5y_0^{(5)} + \cdots.$$

其首项将被用作估计，对变换后的区间，它就变成

$$E_p \approx \frac{14}{45}h^5y_{k-1}^{(5)}.$$

19.19 把预测误差与"校正"公式误差进行比较

$$y_{k+1} \approx y_{k-1} + \frac{1}{3}h(y'_{k-1} + 4y'_k + y'_{k+1}).$$

解 ☞ 这个校正步实际上就是对 $y'(x)$ 用 Simpson 公式,因此根据题 14.65 局部截断误差是

$$E_c = \int_{x_{k-1}}^{x_{k+1}} y'(x)\mathrm{d}x - \frac{1}{3}h(y'_{k-1} + 4y'_k + y'_{k+1}) \approx -\frac{1}{90}h^5 y_k^{(5)}(\xi).$$

因此 $E_p \approx -28E_c$,这里没有顾及 $y^{(5)}$ 的自变量的误差.

19.20 证明能用预测和校正值的差来估计题 19.19 中的校正公式的误差.

证 ☞ 只考虑从 x_k 到 x_{k+1} 步产生的局部截断误差,我们有

$$y_{k+1} = P + E_p = C + E_c,$$

P 和 C 分别表示预测值和校正值,于是

$$P - C = E_c - E_p = 29E_c,$$

$$E_c = \frac{P - C}{29}.$$

经常用这一估计作进一步校正,得到

$$y_{k+1} = C + \frac{P - C}{29}.$$

这个公式确有 h^6 阶的截断误差,但在某些条件下,这些"抹去"(mop-up)项的使用可能使计算不稳定.

19.21 Milne 方法用

$$y_{k+1} \approx y_{k-3} + \frac{4}{3}h(2y'_{k-2} - y'_{k-1} - 2y'_k)$$

作为预测步,同时用

$$y_{k+1} \approx y_{k-1} + \frac{1}{3}h(y'_{k+1} + 4y'_k + y'_{k-1})$$

作为校正步.取 $h = 0.2$ 把这个方法用于 $y' = -xy^2, y(0) = 2$.

解 ☞ 预测步需用前四个值,把它们混合成 y_{k+1},初值 $y(0) = 2$ 是其中之一.必须求得其余三个值.因为全部计算基于这些出发值.所以值得额外地努力使它们适当精确.可使用 Taylor 方法或 Runge-Kutta 方法求出精确到五位小数的

$$y(0.2) = y_1 \approx 1.92308 \quad y(0.4) = y_2 \approx 1.72414 \quad y(0.6) = y_3 \approx 1.47059$$

然而由微分方程得到精确到五位小数的

$$y'(0) = y'_0 = 0, \qquad y'(0.2) = y'_1 \approx -0.73964,$$
$$y'(0.4) = y'_2 \approx -1.18906 \qquad y'(0.6) = y'_3 \approx -1.29758.$$

然后用 Milne 预测步得到

$$y_4 \approx y_0 + \frac{4}{3}(0.2)(2y'_3 - y'_2 + 2y'_1) \approx 1.23056.$$

由微分方程我们得到 y'_4 的首次估计

$$y'_4 \approx -(0.8)(1.23056)^2 \approx -1.21142.$$

再由 Milne 校正步得出新的近似

$$y_4 \approx y_2 + \frac{1}{3}(0.2)(-1.21142 + 4y'_3 + y'_2) \approx 1.21808$$

通过微分方程重新计算 y' 得出新的估计 $y'_4 \approx -1.18698$,其次再用校正步,我们有

$$y_4 \approx y_2 + \frac{1}{3}(0.2)(-1.18698 + 4y'_3 + y'_2) \approx 1.21971.$$

再次利用微分方程,我们得到

$$y'_4 \approx -1.19015.$$

再回到校正步

$$y_4 \approx y_2 + \frac{1}{3}(0.2)(-1.19015 + y'_3 + y'_2) \approx 1.21950.$$

下面两个循环产生

$$y'_4 \approx -1.18974, \qquad y_4 \approx 1.21953, \qquad y'_4 \approx -1.18980 \qquad y_4 \approx 1.21953.$$

因为 y_4 的最后两个估计值相同,所以我们就停止.校正公式和微分方程的迭代使用已证明是收敛

的.结果产生的 y_4 的值精确到四位,在这种情形用四次校正步已经带来收敛性.在这类过程中如果 h 选得太大就可能要过多的迭代,或许算法根本不收敛,预测步和校正步输出之间的大的误差启发我们减小区间.另一方面预测步和校正步输出之间不显著的误差启发我们增大 h 或许能加快计算.现在可以用同样的方法来计算 y_5 和 y'_5.直到 $x=10$ 的结果在表 19.4 中给出.虽然取 $h=0.2$,为简单起见,仅印出整数变量的值,为了进行比较精确值也在其中.

表 19.4

x	y(精确)	y(预测)	误差	y(校正)	误差
0	2.00000				
1	1.00000	1.00037	- 37	1.00012	- 12
2	0.40000	0.39970	30	0.39996	4
3	0.20000	0.20027	- 27	0.20011	- 11
4	0.11765	0.11737	28	0.11750	15
5	0.07692	0.07727	- 35	0.07712	- 20
6	0.05405	0.05364	41	0.05381	14
7	0.04000	0.04048	- 48	0.04030	- 30
8	0.03077	0.03022	55	0.03041	36
9	0.02439	0.02500	- 61	0.02481	- 42
10	0.01980	0.01911	69	0.01931	49

19.22 讨论上述计算的误差.

解 因为这个试验情形中精确解是知道的,容易看到某些事情通常十分模糊的. $y(x)=2/(1+x^2)$的五阶导数所具有的一般性态呈现在图 19.4 中.

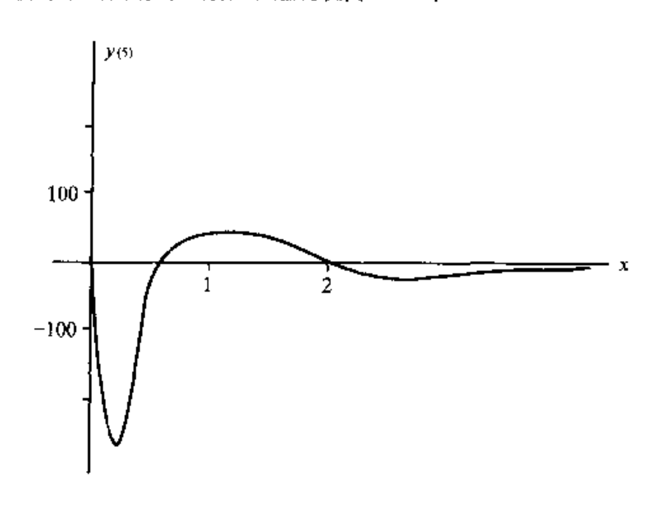


图 19.4

0 和 1 之间的巨大振动使它一般很难用我们的截断误差公式.例如预测步的局部截断误差是 $14h^5y^{(5)}/45$.在我们的第一步(到 $x=0.8$)我们实际求得预测步的误差是 -0.011.这相当 $y^{(5)}\approx-100$.局部校正误差是 $-h^5y^{(5)}/90$,同样在第一步其误差实际上是 -0.00002,这相当于 $y^{(5)}\approx6$.这种 $y^{(5)}$符号的改变宣布预料的预测步和校正步结果之间误差的符号的改变无效.它也说明企图用外推到极限的观念将导致更坏的而不是更好的结果.当继续计算时,误差符号的振荡将在后面讨论.

19.23 导出 Adams 预测公式

$$\begin{aligned}y_{k+1}&=y_k+h\left(y'_k+\frac{1}{2}\nabla y'_k+\frac{5}{12}\nabla^2y'_k+\frac{3}{8}\nabla^3y'_k\right)\\&=y_k+\frac{1}{24}h(55y'_k-59y'_{k-1}+37y'_{k-2}-9y'_{k-3}).\end{aligned}$$

解 ☞ 和在题 19.17 中一样，我们通过在超出配置区间的范围对配置多项式积分得到这个预测步，把三次牛顿向后公式用 $y'(x)$ 即

$$p_k = y'_0 + k\nabla y'_0 + \frac{1}{2}k(k+1)\nabla^2 y'_0 + \frac{1}{6}k(k+1)(k+2)\nabla^3 y'_0,$$

这里 $x_k = x_0 + kh$. 从 $k=0$ 到 $k=1$ 积分(尽管配置点是 $k=0,-1,-2,-3$)我们得到

$$\int_0^1 p_k \mathrm{d}k = y'_0 + \frac{1}{2}\nabla y'_0 + \frac{5}{12}\nabla^2 y'_0 + \frac{3}{8}\nabla^3 y'_0.$$

用自变量 x 表示并利用 $p(x) \approx y'(x)$，它就成为

$$\int_{x_0}^{x_1} y'(x)\mathrm{d}x = y_1 - y_0 \approx h\left(y'_0 + \frac{1}{2}\nabla y'_0 + \frac{5}{12}\nabla^2 y'_0 + \frac{3}{8}\nabla^3 y'_0\right).$$

由于同样的推理可用于 x_k 和 x_{k+1} 之间，所以我们可把所有的下标加上 k，而得到第一个所需的结果. 然后用 y 值写出各个差分就得到第二个式子.

19.24　Adam 预测步的局部截断误差是什么？

解 ☞ 通常的 Taylor 级数逼近导出 $E = 251h^5y^{(5)}/720$.

19.25　导出以下形式的另一种预测步

$$y_{k+1} = a_0y_k + a_1y_{k-1} + a_2y_{k-2} + h(b_0y'_k + b_1y'_{k-1} + b_2y'_{k-2} + b_3y'_{k-3}).$$

解 ☞ 改变这种逼近，我们将使这一公式对直到四次的多项式是准确的. 方便的选择是 $y(x) = 1, (x-x_k), (x-x_k)^2, (x-x_k)^3$ 和 $(x-x_k)^4$，这就导出了五个条件：

$1 = a_0 + a_1 + a_2,$　　　　　　$1 = -a_1 - 8a_2 + 3b_1 + 12b_2 + 27b_3,$

$1 = -a_1 - 2a_2 + b_0 + b_1 + b_2 + b_3,$　　　　$1 = a_1 + 16a_2 - 4b_1 - 32b_2 - 108b_3,$

$1 = a_1 + 4a_2 - 2b_1 - 4b_2 - 6b_3.$

可得到以下形式的解：

$a_0 = 1 - a_1 - a_2,$　　　　　　$b_2 = \frac{1}{24}(37 - 5a_1 + 8a_2),$

$b_0 = \frac{1}{24}(55 + 9a_1 + 8a_2),$　　　　$b_3 = \frac{1}{24}(-9 + a_1),$

$b_1 = \frac{1}{24}(-59 + 19a_1 + 32a_2),$

其中 a_1 和 a_2 任意. 选取 $a_1 = a_2 = 0$ 使我们回到前一问题. 另两个简单而流行的选择是 $a_1 = \frac{1}{2}, a_2 = 0$，它导出

$$y_{k+1} = \frac{1}{2}(y_k + y_{k-1}) + \frac{1}{48}h(119y'_k - 99y'_{k-1} + 69y'_{k-2} - 17y'_{k-3}).$$

其截断误差是 $161h^5y^{(5)}/480$，而 $a_1 = \frac{2}{3}, a_2 = \frac{1}{3}$ 时导出的是

$$y_{k+1} = \frac{1}{3}(2y_{k-1} + y_{k-2}) + \frac{1}{72}h(191y'_k - 107y'_{k-1} + 109y'_{k-2} - 25y'_{k-3})$$

其截断误差是 $707h^5y^{(5)}/2160$.

显然我们能用这两个自由参数进一步去减小截断误差，甚至达到 h^7 阶，但是另一个要简短地加以考虑的因素指出截断误差并非是我们惟一的问题. 也很清楚其他类型的预测步可能用 y_{k-3} 这一项，但是我们将限制我们已有的项的数量.

19.26　说明其他校正公式的可能性.

解 ☞ 可能性很多，但假设我们寻求下面形式的校正步

$$y_{k+1} \approx a_0y_k + a_1y_{k-1} + a_2y_{k-2} + h(cy'_{k+1} + b_0y'_k + b_1y'_{k-1} + b_2y'_{k-2}),$$

使其局部截断误差是 h^5 阶. 要求这个校正步对 $y(x) = 1, (x-x_k), \cdots (x-x_k)^4$ 精确成立就导出五个条件：

$a_0 + a_1 + a_2 = 1,$　　　　　　$13a_1 + 32a_2 - 24b_1 = 5,$

$a_1 + 24c = 9,$　　　　　　$a_1 - 8a_2 + 24b_2 = 1,$

$13a_1 + 8a_2 - 24b_0 = -19,$

其中包括七个未知常数. 如果这一校正步对更高幂次的 x 精确成立，便能进一步减小局部截断误

差.然而这两个自由度,将用来得到其他所要求的特征而不是所产生的算法.当 $a_0 = 0$ 和 $a_1 = 1$ 时,其余的常数证实是 Milne 校正步的那几个常数:

$$a_2 = 0, \qquad c = \frac{1}{3}, \qquad b_0 = \frac{4}{3}, \qquad b_1 = \frac{1}{3}, \qquad b_2 = 0.$$

和某种推广的 Adams 预测步匹配的另一种选择即取 $a_1 = a_2 = 0$,就得到以下公式

$$y_{k+1} \approx y_k + \frac{1}{24}h(9y'_{k+1} + 19y'_k - 5y'_{k-1} + y'_{k-2}).$$

假如 $a_1 = \frac{2}{3}, a_2 = \frac{1}{3}$,那么我们就得到类似于刚才说到的另一个预测公式

$$y_{k+1} \approx \frac{1}{3}(2y_{k-1} + y_{k-2}) + \frac{1}{72}h(25y'_{k+1} + 91y'_k + 43y'_{k-1} + 9y'_{k-2}).$$

还有另一种公式取 $a_0 = a_1 = \frac{1}{2}$ 得到

$$y_{k+1} \approx \frac{1}{2}(y_k + y_{k-1}) + \frac{1}{48}h(17y'_{k+1} + 51y'_k + 3y'_{k-1} + y'_{k-2}).$$

不同的选择使它们的截断误差多少有些不同.

19.27　比较上述预测和校正公式的局部截断误差.

解　☞ 仍能用 Taylor 级数方法得到以下的误差估计:

预测:　$y_{k+1} = y_k + \frac{1}{24}h(55y'_k - 59y'_{k-1} + 37y'_{k-2} - 9y'_{k-3}) + \frac{251h^5y^{(5)}}{720}$.

校正:　$y_{k+1} = y_k + \frac{1}{24}h(9y'_{k+1} + 19y'_k - 5y'_{k-1} + y'_{k-2}) - \frac{19h^5y^{(5)}}{720}$.

预测:　$y_{k+1} = \frac{1}{2}(y_k + y_{k-1}) + \frac{1}{48}h(119y'_k - 99y'_{k-1} + 69y'_{k-2} - 17y'_{k-3}) + \frac{161h^5y^{(5)}}{480}$.

校正:　$y_{k+1} = \frac{1}{2}(y_k + y_{k-1}) + \frac{1}{48}h(17y'_{k+1} + 51y'_k + 3y'_{k-1} + y'_{k-2}) - \frac{9h^5y^{(5)}}{480}$.

预测:　$y_{k+1} = \frac{1}{3}(2y_{k-1} + y_{k-2}) + \frac{1}{72}h(191y'_k - 107y'_{k-1} + 109y'_{k-2} - 25y'_{k-3}) + \frac{707h^5y^{(5)}}{2160}$.

校正:　$y_{k+1} = \frac{1}{3}(2y_{k-1} + y_{k-2}) + \frac{1}{72}h(25y'_{k+1} + 91y'_k + 43y'_{k-1} + 9y'_{k-2}) - \frac{43h^5y^{(5)}}{2160}$.

在每一种情形校正误差比与它搭配的预测误差要小得多.它还带有相反的符号.这在计算中可能是一个有益的信息.较低的校正误差能用其由来而得到解释,它用了有关 y'_{k+1} 的信息,而预测公式必须取自 y_k,这也说明为什么计算的负担落在校正步而预测步仅用作一个初步.

对每一组公式都能推出抹掉的项.上面的第一组取 Adams 预测及校正在其下.以通常的方法进行下去,只考虑局部截断误差并且仍认为这样得到的结果值得怀疑.我们得到

$$I = P + E_1 = C + E_2.$$

其中 I 是精确值.因为 $19E_1 \approx -251E_2$,所以我们有 $E_2 \approx \frac{19}{270}(P - C)$,这就是抹掉项,而 $I \approx C + \frac{19}{270}(P - C)$ 相应于外推到极限.必须再次记住在这两个公式中 $y^{(5)}$ 其实并不表示同一个东西*,所以在这一外推中仍有可能存在相当大的误差.

19.28　取 $h = 0.2$,对 $y' = -xy^2, y(0) = 2$ 应用 Adams 方法.

解　☞ 现在这个方法已经熟悉,每一步包括预测,然后是校正公式的反复使用.Adams 方法用题 19.27 中的第一组公式导出了表 19.5 的结果.

* 译注:意指它们不一定是在同一点上取的值.

表 19.5

x	y(正确)	y(预测)	误差	y(校正)	误差
0	2.000000				
1	1.000000	1.000798	−789	1.000133	−133
2	0.400000	0.400203	−203	0.400158	−158
3	0.200000	0.200140	−140	0.200028	−28
4	0.117647	0.117679	−32	0.117653	−6
5	0.076923	0.076933	−10	0.076925	−2
6	0.054054	0.054058	−4	0.054055	−1
7	0.040000	0.040002	−2	0.040000	
8	0.030769	0.030770	−1	0.030769	
9	0.024390	0.024391	−1	0.024390	
10	0.019802	0.019802		0.019802	

误差的性态显示对于大的 x, $h=0.2$ 足以得到六位精度. 但是在开始时, 取较小的 h(如 0.1)可能是合理的, 缩小误差关系到这个事实(见题 19.36)即这种方法的"相对误差"保持有界.

19.29 对充分小的 h, 证明反复使用校正公式产生一个收敛序列, 而且这一序列的极限是满足校正公式的惟一值 Y_{k+1}.

解 我们求一个数 Y_{k+1}满足性质

$$Y_{k+1}=hcf(x_{k+1},Y_{k+1})+\cdots.$$

省略号表示仅含前面已算得的结果的项, 因此与 Y_{k+1}无关. 和通常一样, 假设在区域 R 内 $f(x,y)$ 对 y 满足 Lipschitz 条件. 现在定义序列

$$Y^{(0)},\ Y^{(1)},\ Y^{(2)}.$$

为简单计, 省去下标 $k+1$. 借助迭代公式

$$Y^{(i)}=hcf(x_{k+1},Y^{(i-1)})+\cdots,$$

并假设所有的点$(x_{k+1},Y^{(i)})$在 R 内. 相减得

$$Y^{(i+1)}-Y^{(i)}=hc[f(x_{k+1},Y^{(i)})-f(x_{k+1},Y^{(i-1)})].$$

重复用 Lipschitz 条件就得到

$$|Y^{(i+1)}-Y^{(i)}|\leqslant hcK|Y^{(i)}-Y^{(i-1)}|\leqslant\cdots\leqslant(hcK)^i|Y^{(1)}-Y^{(0)}|.$$

现选取 h 足够小使得, $|hcK|<1$, 同时考虑和

$$Y^{(n)}-Y^{(0)}=(Y^{(1)}-Y^{(0)})+\cdots+(Y^{(n)}-Y^{(n-1)}),$$

当 n 趋于无穷时, 右端所得到的级数被几何级数 $1+r+r^2+\cdots$所控制(相差一个因子), 因此收敛. 这就证明了 $Y^{(n)}$存在极限, 称此极限为 Y_{k+1}. 现在根据 Lipschitz 条件

$$|f(x_{k+1},Y^{(n)})-f(x_{k+1},Y_{k+1})|\leqslant K|Y^{(n)}-Y_{k+1}|.$$

因此 $\lim f(x_{k+1},Y^{(n)})=f(x_{k+1},Y_{k+1})$. 于是我们可以在迭代

$$Y^{(n)}=hcf(x_{k+1},Y^{(n-1)})+\cdots$$

中令 n 趋于无穷, 就立即得到所需要的

$$Y_{k+1}=hcf(x_{k+1},Y_{k+1})+\cdots.$$

为了证明惟一性, 假设 Z_{k+1}是在 x_{k+1}处满足校正公式的另一值. 类似前面一样因为

$$|Y_{k+1}-Z_{k+1}|\leqslant hcK|Y_{k+1}-Z_{k+1}|\leqslant\cdots\leqslant(hcK)^{(i)}|Y_{k+1}-Z_{k+1}|,$$

对任意 i 都成立. 而$|hcK|=r<1$, 所以必有 $Y_{k+1}=Z_{k+1}$. 注意这个惟一性结论说明了校正值 Y_{k+1}与 $Y^{(0)}$无关. 即至少对小的 h 与预测公式的选取无关. 因此选取预测公式十分自由. 对一个给定的校正公式, 从局部截断误差的观点考虑, 使用较精确的预测公式是合理的. 这又导出了吸引人的"抹掉(mop-up)"量. 记住题 19.27 中的各种组合保持了这些因子而且是一些简单的漂亮因子.

预测-校正方法的收敛性

19.30 证明修正的 Euler 方法是收敛的.

证 在这个方法中, 用简单 Euler 公式对每个 y_{k+1}作首次预测, 但是实际近似是用修正的

Euler 公式

$$Y_{k+1} = Y_k + \frac{1}{2}h(Y'_{k+1} + Y'_k)$$

得到的. 精确解满足带有截断误差项的类似关系式, 和前面一样, 称精确解为 $y(x)$, 我们有

$$y(x_{k+1}) = y(x_k) + \frac{1}{2}h[y'(x_{k+1}) + y'(x_k)] - \frac{1}{12}h^3 y^{(3)}(\xi).$$

截断误差项已在题 19.15 中作了估计. 相减, 并令 $d_k = y(x_k) - Y_k$, 就有

$$|d_{k+1}| \leqslant |d_k| + \frac{1}{2}hL(|d_{k+1}| + |d_k|) + \frac{1}{12}h^3 B.$$

只要我们假定满足 Lipschitz 条件, 即

$$|y'(x_k) - Y'_k| = |f(x_k, y(x_k)) - f(x_k, Y_k)| \leqslant L|d_k|.$$

上式对 $k+1$ 同样成立. 数 B 是 $|y^{(3)}(x)|$ 的界, 我们已假定它是存在的. 我们的不等式也可写为

$$\left(1 - \frac{1}{2}hL\right)|d_{k+1}| \leqslant \left(1 + \frac{1}{2}hL\right)|d_k| + \frac{1}{12}h^3 B.$$

假设没有初始误差 ($d_0 = 0$), 同时考虑带初值 $D_0 = 0$

$$\left(1 - \frac{1}{2}hL\right)D_{k+1} = \left(1 + \frac{1}{2}hL\right)D_k + \frac{1}{12}h^3 B$$

的解. 为了进行递推, 我们假设 $|d_k| \leqslant D_k$, 因此

$$\left(1 - \frac{1}{2}hL\right)|d_{k+1}| \leqslant \left(1 - \frac{1}{2}hL\right)D_{k+1}.$$

所以 $|d_{k+1}| \leqslant D_{k+1}$, 因为 $d_0 = D_0$ 所以完成了归纳法. 这就保证了对任何整数 k 都成立 $|d_k| \leqslant D_k$. 为了求 D_k, 我们解差分方程并求得解族

$$D_k = C\left(\frac{1 + \frac{1}{2}hL}{1 - \frac{1}{2}hL}\right)^k - \frac{h^2 B}{12L},$$

其中 C 是任意常数. 为了满足初始条件 $D_0 = 0$, 我们必须有 $C = (h^2 B/12L)$, 所以

$$|y(x_k) - Y_k| \leqslant \frac{h^2 B}{12L}\left[\left(\frac{1 + \frac{1}{2}hL}{1 - \frac{1}{2}hL}\right)^k - 1\right].$$

为了证明在固定的 $x_k = x_0 + kh$ 处收敛, 我们研究第二个因子. 这是因为当 h 趋于零时, k 无限增大. 但是因为

$$\left(\frac{1 + \frac{1}{2}hL}{1 - \frac{1}{2}hL}\right)^k = \left[\frac{1 + L(x_k - x_0)/2k}{1 - L(x_k - x_0)/2k}\right]^k \to \frac{e^{L(x_k - x_0)/2}}{e^{-L(x_k - x_0)/2}} = e^{L(x_k - x_0)},$$

我们有
$$y(x_k) - Y_k = O(h^2).$$

因此当 h 趋于零时, $\lim Y_k = y(x_k)$, 即收敛. 我们的结果也提供了在计算过程中截断误差传播方式的度量.

19.31 证明 Milne 方法的收敛性.

证 Milne 校正公式本质上是 Simpson 法则, 并提供近似值

$$Y_{k+1} = Y_{k-1} + \frac{1}{3}h(Y'_{k+1} + 4Y'_k + Y'_{k-1}).$$

精确解 $y(x)$ 满足类似的关系式, 但带有截断误差项

$$y_{k+1} = y_{k-1} + \frac{1}{3}h(y'_{k+1} + 4y'_k + y'_{k-1}) - \frac{1}{90}h^5 y^{(5)}(\xi) \qquad (\text{注}: y_k = y(x_k)),$$

这里 ξ 在 x_{k-1} 和 x_{k+1} 之间. 相减并令 $d_k = y(x_k) - Y_k$,

$$|d_{k+1}| \leqslant |d_{k-1}| + \frac{1}{3}hL(|d_{k+1}| + 4|d_k| + |d_{k-1}|) + \frac{1}{90}h^5 B.$$

这里又一次用到了 Lipschitz 条件及 $y^{(5)}(x)$ 的上界 B. 将等式改写为

$$\left(1 - \frac{1}{3}hL\right)|d_{k+1}| \leqslant \frac{4}{3}hLd_k + \left(1 + \frac{1}{3}hL\right)|d_{k-1}| + \frac{1}{90}h^5 B.$$

我们把它与差分方程

$$\left(1-\frac{1}{3}hL\right)D_{k+1}=\frac{4}{3}hLD_k+\left(1+\frac{1}{3}hL\right)D_{k-1}+\frac{1}{90}h^5B$$

进行比较.

假设初始误差是 d_0 和 d_1. 我们求解 D_k 使得 $d_0\leqslant D_0$ 和 $d_1\leqslant D_1$. 这个解将控制 $|d_k|$, 即对非负整数 k 成立 $|d_k|\leqslant D_k$. 这可以用归纳法证明. 与前题十分类似, 假设 $|d_{k-1}|\leqslant D_{k-1}$ 及 $|d_k|\leqslant D_k$, 我们可以立刻得到 $|d_{k+1}|\leqslant D_{k+1}$ 即归纳法已完成. 为了求出所需的解, 可以先解特征方程

$$\left(1-\frac{1}{3}hL\right)r^2-\frac{4}{3}hLr-\left(1+\frac{1}{3}hL\right)=0.$$

容易发现一个根略大于 1, 譬如说 r_1, 而另一个根在 -1 附近, 譬如说 r_2. 更准确地说,

$$r_1=1+hL+0(h^2),\qquad r_2=-1+\frac{1}{3}hL+0(h^2),$$

相关联的齐次方程的解是这些根的 k 次幂的组合, 非齐次方程本身有常数解 $-h^4B/180L$. 因此我们有

$$D_k=c_1r_1^k+c_2r_2^k-\frac{h^4B}{180L}.$$

设 E 是两数 d_0 和 d_1 中间较大的一个, 则

$$D_k=\left(E+\frac{h^4B}{180L}\right)r_1^k-\frac{h^4B}{180L}$$

就是满足初始条件的解. 由上式可得 $D_0=E$. 并且因为 $1<r_1$, 所以 D_k 不断增大. 于是

$$|d_k|\leqslant\left(E+\frac{h^4B}{180L}\right)r_1^k-\frac{h^4B}{180L}.$$

如果我们没有初始误差, 于是 $d_0=0$. 另外, 如果当 h 变小, 我们改进 Y_1 值(这须用某些其他方法, 譬如 Taylor 级数得到)使得 $d_1=0(h)$, 于是就有 $E=0(h)$, 并且当 h 趋于零, d_k 也趋于零. 这就证明了 Milne 方法的收敛性.

19.32　推广前题, 证明基于校正公式

$$Y_{k+1}=a_0Y_k+a_1Y_{k-1}+a_2Y_{k-2}+h(cY'_{k+1}+b_0Y'_k+b_1Y'_{k-1}+b_2Y'_{k-2})$$

的方法的收敛性.

证　☞ 我们已选取适当的系数使得截断误差为 h^5 阶. 假设就是这种情形, 恰如对 Milne 校正公式采用的相同的步骤, 发现差 $d_k=y(x_k)-Y_k$ 满足

$$(1-|c|hL)|d_{k+1}|\leqslant\sum_{i=0}^{2}(|a_i|+hL|b_i|)|d_{k-i}|+T,$$

这里 T 是截断误差项. 这个校正公式需要三个出发值, 也可以由 Taylor 级数求得令这些值的最大误差为 E, 于是 $|d_k|\leqslant E$, $k=0,1,2$, 同时考虑差分方程

$$(1-|c|hL)D_{k+1}=\sum_{i=0}^{2}(|a_i|+hL|b_i|)D_{k-i}+T,$$

我们对 $k=0,1,2$ 求一个解满足 $E\leqslant D_k$. 这样的解将控制 $|d_k|$. 因为假定对 $i=0,1,2$ 成立 $|d_{k-i}|\leqslant D_{k-i}$, 我们立刻有 $|d_{k+1}|\leqslant D_{k+1}$, 这就完成了归纳法. 即对非负整数 k 证明了 $|d_k|\leqslant D_k$. 为了求得所需要的解, 我们注意特征方程

$$(1-|c|hL)r^3-\sum_{i=0}^{2}(|a_i|+hL|b_i|)r^{2-i}=0$$

有一实根大于 1. 这是因为在 $r=1$ 时, 其左端成为

$$A=1-|c|hL-\sum_{i=0}^{2}(|a_i|+hL|b_i|),$$

因为 $a_0+a_1+a_2=1$, 所以 A 必为负, 而对于大的 r, 如果我们选取 h 充分小使 $1-|c|hL$ 为正, 那么其左端必为正. 称特征方程的这个根为 r_1, 则满足要求的解为

$$D_k=\left(E-\frac{T}{A}\right)r_1^k+\frac{T}{A}.$$

由于在 $k=0$ 时, 它等于 E, 而当 k 增加时它变得更大, 因此

$$|y(x_k)-Y_k|\leqslant\left(E-\frac{T}{A}\right)r_1^k+\frac{T}{A}.$$

当 h 趋于零时, 截断误差 T 趋于零. 如果我们还调正初始误差趋于零, 那么 $\lim y(x_k)=Y_k$, 即收敛

性得证.

误差和稳定性

19.33 用稳定方法求解微分方程是什么意思?

解 ☞ 稳定的概念已用很多方法加以描述.很不严格地说,如果一种计算不会弄糟,它就是稳定的.但是这几乎不能作为正式的定义.在本章的引言中稳定性定义是定义为相对误差有界.毫无疑问,对于一个算法来说这将是所希望的特征.相对误差的逐渐增大表示有效数字的逐渐丧失,这不是我们所期盼的.麻烦的是经过长期运算,常使相对误差增加.能用一个简单的例子来说明.考虑改进的 Euler 方法.

$$y_{k+1}=y_k+\frac{1}{2}h(y'_{k+1}+y'_k),$$

把它用于一般问题 $y'=Ay,\quad y(0)=1.$

它的精确解是 $y=e^{Ax}$.Euler 公式变为

$$\left(1-\frac{1}{2}Ah\right)y_{k+1}=\left(1+\frac{1}{2}Ah\right)y_k.$$

这是一阶差分方程,它的解是

$$y_k=r^k=\left(\frac{1+\frac{1}{2}Ah}{1-\frac{1}{2}Ah}\right)^k.$$

对于小的 h,它接近于 $\left(\frac{e^{(1/2)Ah}}{e^{-(1/2)Ah}}\right)^k=e^{Akh}=e^{Ax}$,

这就给我们提供了直接的收敛性证明.但是我们这里的目标不在这方面.精确解满足

$$\left(1-\frac{1}{2}Ah\right)y(x_{k+1})=\left(1+\frac{1}{2}Ah\right)y(x_k)+T.$$

这里 T 是截断误差 $-h^3A^3y(\xi)/12$,相减并令 $d_k=y(x_k)-y_k$,对 d_k 得类似的方程

$$\left(1-\frac{1}{2}Ah\right)d_{k+1}=\left(1+\frac{1}{2}Ah\right)d_k-\frac{1}{12}h^3A^3y(\xi).$$

除以 $\left(1-\frac{1}{2}Ah\right)y(x_{k+1})$*,并假设 Ah 很小,就得到,对于相对误差

$$R_k=d_k/y_{x_k}{}^{*}\qquad 成立$$

$$R_{k+1}=R_k-\frac{1}{12}h^3A^3$$

由此可解得 $R_k=R_0-\frac{1}{2}kh^3A^3=R_0-\frac{1}{2}x_kh^2A^3$

这说明在进行计算时,相对误差和 x_k 一样增长或是随 x_k 线性地增长.这虽然没有弄糟,但是也不是使相对误差保持有界的情形.

从另一观点看,我们将观察贯穿求解过程中单个误差的增长.例如对于初始误差 d_0,假定没有其他误差,我们略去 T,就得到

$$d_k=d_0\left(\frac{1+\frac{1}{2}Ah}{1-\frac{1}{2}Ah}\right)^k\approx d_0e^{Akh},$$

这就使得相对误差 $R_k=d_k/e^{Akh}\approx d_0$.所以任一单个误差的长期效应是解本身性态的模拟.若 A 为正,则误差和解以相同的比例增加,若 A 为负,则它们以相同的比例衰减.在这两种情形相对误差都保持不变.这种观点略优于上面预言的线性增长.但是预报了起码不会弄糟.按照某些定义这就足以认为 Euler 算法是稳定的.这种非正规、不严格地使用这个术语可能很方便.

还存在的问题是 Ah 取多小才证实这里的近似是合理的.因为真解是单调的,看来使 $\left(1+\frac{1}{2}Ah\right)/\left(1-\frac{1}{2}Ah\right)$ 的值取正的,是合理的.仅当 Ah 在 -2 和 2 之间时,这才成立.为谨慎起见应使它与两个端点保持一定的距离.

* 译注:原文有错.

19.34 分析 Milne 校正公式

$$y_{k+1}=y_{k-1}+\frac{h}{3}(y'_{k+1}+4y'_k+y'_{k-1})$$

的误差.

解 ☞ 仍然选择特殊的方程 $y'=Ay$,容易发现误差 d_k 满足二阶差分方程

$$\left(1-\frac{1}{3}Ah\right)d_{k+1}=\frac{4}{3}Ahd_k+\left(1+\frac{1}{3}Ah\right)d_{k-1}+T.$$

它的特征方程是(见十八章)

$$\left(1-\frac{1}{3}Ah\right)r^2-\frac{4}{3}Ahr-\left(1+\frac{1}{3}Ah\right)=0.$$

根是 $\qquad r_1=1+Ah+0(h^2),\qquad r_2=-1+\frac{1}{3}Ah+0(h^2).$

这就使 $\qquad d_k\approx c_1(1+Ah)^k+c_2\left(-1+\frac{1}{3}Ah\right)^k$

$$\approx c_1e^{Ahk}+(d_0-c_1)(-1)^k e^{-Ahk/3}.$$

现在就能看到初始误差 d_0 的长期效应,若 A 为正,由于第二项趋于零,所以 d_k 的性态很像准确解 e^{Ahk},事实上相对误差能被估计为

$$\frac{d_k}{e^{Ahk}}=c_1+(d_0-c_1)(-1)^k e^{-4Ahk/3},$$

它趋于常数.但是若 A 为负,第二项不再消失.事实上它很快成为控制项.相对误差呈无界振荡.因此超过某点计算就退化为没有意义.

Milne 方法当 A 是正数时是稳定的,当 A 是负数时是不稳定的.在第二种情形,计算所得的解确实弄糟了.

19.35 前面所作计算是否证实这些理论预测?

解 ☞ 再次参照表 19.4 可以算得以下相对误差.尽管方程 $y'=-xy^2$ 不是线性的.但它的解是不降的,这就像线性方程的解对负数 A 表现的那样.在以上数据中振荡是明显的.相对误差实质性地增长也是明显的.

x_k	1	2	3	4	5	6	7	8	9	10
d_k/y_k	−0.0001	0.0001	−0.0005	0.0013	−0.0026	0.0026	−0.0075	0.0117	−0.0172	0.0247

19.36 分析 Adams 校正公式

$$Y_{k+1}=Y_k+\frac{1}{24}h(9Y'_{k+1}+19Y'_k-5Y'_{k-1}+Y'_{k-2})$$

的误差性态.

解 ☞ 在这种情形,通常的进程能导出

$$\left(1-\frac{9}{24}Ah\right)d_{k+1}=\left(1+\frac{19}{24}Ah\right)d_k-\frac{5}{24}Ahd_{k-1}+\frac{1}{24}Ahd_{k-2}+T$$

不顾及 T,我们企图发现单个误差是如何传播的,特别是在长期运算后,它对相对误差有什么影响?

第一步是再次考虑特征方程

$$\left(1-\frac{9}{24}Ah\right)r^3-\left(1+\frac{19}{24}Ah\right)r^2+\frac{5}{24}Ahr-\frac{1}{24}Ah=0$$

的根.它有一个根接近 1.可以证明 $r_1\approx 1+Ah$,若提出这个根后,留下一个二次因式

$$(24-9Ah)r^2-4Ahr+Ah=0.$$

若 $Ah=0$,这个二次式就有两个零根.若 Ah 不为零,但它很小,这两个根(称为 r_1 和 r_2)仍将接近零.实际上,对小的正数 Ah,它们是模 $|r|\approx\sqrt{Ah/24}$ 的复数.而对小的负数 Ah,它们是实数,而且近似等于 $\pm\sqrt{-6Ah}/12$.同样对小的 Ah,我们有

$$|r_2|,|r_3|<1+Ah\approx e^{Ah}.$$

差分方程的解现在可写为

$$d_k \approx c_1(1+Ah)^k + 0(|Ah|^{k/2}) \approx c_1 e^{Akh} + 0(e^{Akh}).$$

常数 c_1 与已假设的个别误差有关,除以精确解,我们发现相对误差保持有界,因此 Adams 校正公式对正的和负的 A 都是稳定的,单个误差没有破坏计算.

19.37 前面所作计算是否证实这些理论预测?

解 ☞ 再次参照表 19.5,可以算得以下相对误差.

x_k	1	2	3	4	5	6	7 至 10
d_k/y_k	-0.00013	-0.00040	-0.00014	-0.00005	-0.00003	-0.00002	0

如预测的那样,误差减少,甚至相对误差也是这样.再次证明从线性问题得到的结论,对非线性问题计算的性态是有益的.

19.38 寄生解(parasitic solutions)是什么? 它与作为前面问题基础的计算稳定性有什么联系?

解 ☞ 问题的方法包括用差分方程代替微分方程,对于 $y'=Ay$ 这一情形,这就是一个常系数线性差分方程,因此它的解的形式是 r_i^k 项的组合,其中 r_i 是特征方程的根.其中一个根是 $r_1=1+Ah$,并不是 h 的高次项.因此当 h 是小量时,r_1^k 接近 $e^{Akh}=e^{Ax}$,这就是我们想要的解,这个解收敛于微分方程的解.相应于别的 r_i 的其他分量称为再生解.它们是为了使诸如 Milne 方法、Adams 方法有较低的截断误差所付的代价.如果再生解被 r_1 项所控制,那么它的贡献将被忽略,其相对误差将被接受.另一方面,若再生解成为控制项,它就会破坏这一计算.在题 19.33 中,对于修正的 Euler 方法,其相应的差分方程(特征方程)只有根

$$r_1=\frac{1+Ah/2}{1-Ah/2}=1+Ah+0(h^2).$$

就没有再生解.在题 19.34 中 Milne 方法提供给我们

$$r_1=1+Ah,\qquad r_2=-1+\frac{1}{3}Ah$$

直到 h^2.当 $A>0$ 时,由 r_1 控制,但是当 $A<0$ 时,由 r_2 接替而葬送了所要求的解.在题 19.36 中,除了通常的 $r_1=1+Ah$,我们求得两个寄生解项,大小都约为 Ah,不论 A 是正或负,两者都被 r_1 所控制.在这种情形 Adams 方法即为稳定算法.我们得出结论,为了避免计算弄糟,任何再生项都应被主项所控制,即我们要求

$$|r_i|\leqslant r_1,$$

对 $i\neq 1$ 成立.违反这些条件的任何方法都称为不稳定的.事实上,这个不等式最好在更大的范围内成立.

19.39 把二阶 Runge-Kutta 方法

$$y_{k+1}=y_k+hf\left(x_k+\frac{1}{2}h,y_k+\frac{1}{2}hf(x_k,y_k)\right)$$

用于 $y'=Ay, y(0)=1$*.这个公式的稳定性揭示了什么?

解 ☞ 用 Ay 代替 $f(x,y)$ 得到

$$y_{k+1}=\left(1+Ah+\frac{1}{2}A^2h^2\right)y_k.$$

因此

$$y_k=\left(1+Ah+\frac{1}{2}A^2h^2\right)^k.$$

如果 Ay 是小量,它就接近真解 $y(x_k)=e^{Akh}=e^{Ax_k}$**.但是 Ah 应多小? 图 19.5 给出了二次曲线 $r=1+Ah+\frac{1}{2}A^2h^2$ 的图形.当 A 是正数时,r 大于 1,因此 r^k 和 e^{Akh} 都增长,因此 r^k 的性态是正确

* 译注:原文无.

** 译注:原文错为 $y_k=e^{kh}$.

的.但当 A 是负数时,我们要的是下降的解,这仅当 Ah 在 -2 和 0 之间才出现.而在这个区间内近似解可能是增加的,因此与 e^{Akh} 不再相似.因为 Runge-Kutta 方法计算时不到达 y_k 之后,所以这里不存在寄生解.相对误差的"淹没"有不同的原因即在于根 r_1 本身的性质.

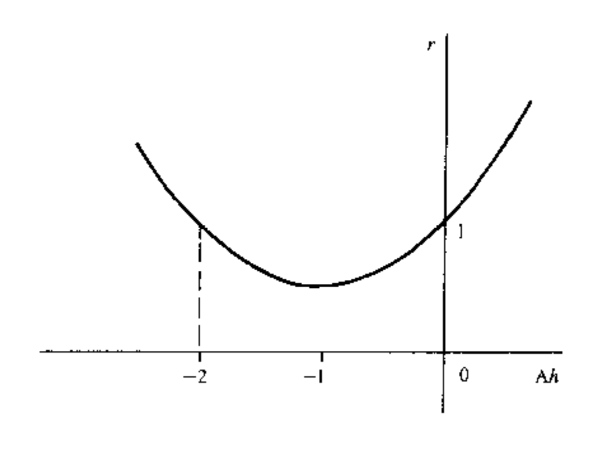


图 19.5

19.40　把题 19.12 中的四阶 Runge-Kutta 公式用到 $y'=Ay$ 上,Ah 的值在什么范围内它稳定?

解　☞ 我们仔细地求得

$$y_{k+1}=\left(1+Ah+\frac{1}{2}A^2h^2+\frac{1}{6}A^3h^3+\frac{1}{24}A^4h^4\right)y_k.$$

这个近似解趋于 e^{Akh} 是明显的,用 r 表示它,我们的近似解又是 $y_k=r^k$;图 19.6 给出了 r 关于 Ah 的曲线图.和二阶方法一样,告诉我们对于正的 A,真解和近似解有相同的特性即都是持续增长的.但是对于负的 A,和前面题一样,存在一个区间下界,在这个区间内 r^k 将不同于真解下降的趋势,此处这个下界靠近 -2.78.对于比这个值小的 Ah,我们发现 r 大于 1,并且破坏了计算.

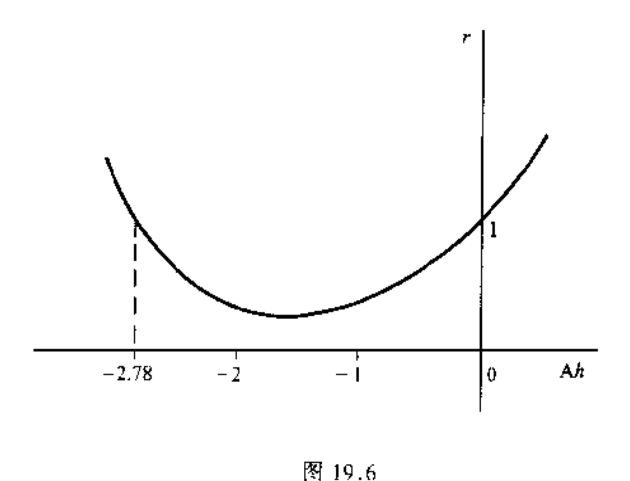


图 19.6

19.41　怎样才能使基于方程 $y'=Ay$ 的分析告诉我们一切都对一般问题 $y'=f(x,y)$ 是有用的?

解　☞ 的确不能够保证,但是对一般方程作这样的分析太困难了,因此真正的问题是哪些是可能做的.在这两个问题间能建立的联系是常数 A 等同于偏导数 f_y,它开始在初始点 (x_0,y_0) 附近,随后是在解所通过平面的其他区域求值.如果 f_y 沿这方向改变符号,我们就预期 Milne 方法的稳定性立即受到影响.Runge-Kutta 方法的稳定性也将显出某种敏感性.

19.42　把四阶 Runge-Kutta 方法用于非线性方程 $y'=-100xy^2$, $y(0)=2$,它的精确解是 $y=2/(1+100x^2)$.对不同步长试验稳定性.

解 ☞ 因为 $f_y=-200xy=-400x/(1+100x^2)$，它开始时等于零，但在 $x=0.1$ 时很快爬到 -20，我们回想起稳定性条件

$$-2.78 \leqslant Ah = -20h$$

并确定在 0.14 周围试验 h 值. 取 $h=0.10$ 所算得的解在 $x=1$ 和 $x=2$ 处刚好衰减到 0.0197 和 0.0050. 取 $h=0.12$ 可观察到类似的下降. 但是取 $h=0.13$，三步就给我们一个很不满意的值 29.11，随后就溢出. 这种肯定的"淹没"对把线性稳定性的准则转到非线性情形作了很好的说明.

19.43 能做什么来控制舍入误差?

解 ☞ 在很长的求解过程中，舍入误差可能成为很重要的因素. 如果双精度算法可以用的话，它可能就是应该用的，而不论额外的浪费，即只能求助它了. 假定认为整个计算中使用高精度太耗费时间，有中间步可能很有帮助. 为了说明这一点，许多解微分方程的公式相当于

$$y_{k+1} = y_k + h\Delta y_k.$$

与 y_k 本身相比，Δy_k 是小量，执行右端加法时，这个小的校正项移位(对准二进制小数点)，就在这里产生了舍入误差. 为避免这种舍入误差，把 y_k 以双精度贮存并以双精度进行这种加法. 计算 Δy_k 的工作通常是最繁重的工作，仍然用单精度计算 Δy_k，这是因为这一项反正预计是小量. 在这个方法中双精度仅用于最需要的地方.

自适应方法，变步长

19.44 如何把题 14.27 中介绍的自适应积分的概念推广到处理微分方程?

解 ☞ 假设目标是从起点 $x=a$ 到终点 $x=b$ 近似求解 $y'=f(x,y)$，误差不超过 e. 设误差线性累积使得每经过长度为 h 的一步后，我们能允许大小为 $eh/(b-a)$ 的误差. 这恰好是以前用过的自适应积分. 假设 T 是取长度为 h 的一步所产生的截断误差估计. 因此如果 T 不超过 $eh/(b-a)$，就采用这一步，并继续进行下一步. 否则，减小步长 h(到 $0.5h$ 或其他适合的)，并重复这一步. 只要步长不是小得让舍入误差成为控制误差源，用收敛方法这些要求最终会被满足.

如果用 Milne 预测-校正方法，那么题 19.20 提供了需要的截断误差 $(P-C)/29$，能接受的条件是

$$|P-C| \leqslant \frac{29eh}{b-a}$$

它容易从已处理的部分计算出来.

如果用 Adams 方法，那么题 19.27 给出同样能接受的条件是

$$|P-C| \leqslant \frac{\frac{270}{19}eh}{b-a}$$

在这种情形，舍弃将需要重新使用附加的启动程序.

19.45 为了使 Runge-Kutta 方法自适应，需要一种估计局部截断误差的切实可行的方法. 建立这种估计，即不含 $y(x)$ 高阶导数的那种.

解 ☞ 现在将用到熟悉的对步长 h 和 $2h$ 的误差进行比较的概念. 用经典的四阶方法，在目前的位置 x_k 取步长为 $2h$，局部截断误差大约为

$$T_{2h} = C(2h)^5 = 32Ch^5.$$

现在用步长为 h 的两步来覆盖同样的区间，合起来的误差约为

$$2T_h = 2Ch^5.$$

这就是对真的 y_{k+2} 值的两个估计：

$$y_{k+2} = A_{2h} + 32Ch^5 = A_h + 2Ch^5,$$

下标 $2h$ 和 h 表示得到这两种近似用的步长，相减就得到 C 的值和误差估计

$$T_h = Ch^5 = \frac{A_h - A_{2h}}{30}.$$

完整的向前步可能是它的两倍，这种估计假定 Ch^5 是一个合适的误差度量而 C(有高阶导数夹在中间)在整个区间改变不大.

19.46 利用前一题的误差估计，使 Runge-Kutta 方法自适应.

解 ☞ 对于区间(a,b)允许误差是 e，为了使它按比例分布，我们要求在 x_k 和 x_{k+1}之间的局部截断误差不超过 $2eh/(b-a)$. 如果 T_{2h}恰如估计的不超过此值，即若

$$|A_h - A_{2h}| \leqslant \frac{30eh}{b-a},$$

那么 A_h 的值能在x_{k+2}处被接受，而且可以继续下去. 换言之，为了使新的截断误差 T_{h^*}是合适的就需要更小的步长 h^*. 回到基本的，我们假设

$$T_h = Ch^5, \qquad T_{h^*} = Ch^{*5} = \frac{T_h h^{*5}}{h^5},$$

后者数量上不超过 $h^* e/(b-a)$，放在一起就得到新步长是

$$h^* = \left[\frac{eh^5}{(b-a)T_h}\right]^{1/4}.$$

从导出这公式的各种假设看来，它并没有走向极端，通常引入保险因子 0.8. 此外假设 h 已经很小，T_h 随之而很小，h^* 的计算甚至可能溢出. 此公式应慎重地应用.

19.47 预测校正方法和 Runge-Kutta 方法中哪一种更适合于自适应计算？

解 ☞ 预测-校正方法的优点是，估计截断误差的部分当需要时已经处理，对于 Runge-Kutta 方法必须分开使用公式，恰如刚才提到的. 这几乎使必须计算 $f(x,y)$的时间加倍，因为这是所含的主要计算工作，所以运算时间几乎加倍. 另一方面，如前所说每当步长改变，就需要帮助预测-校正方法重新启动，这就意味着额外的程序. 因此若预知经常改变，不如整个使用 Runge-Kutta 方法.

19.48 试改变步长用经典 Runge-Kutta 方法解问题

$$y' = -xy^2, \qquad y(0) = 2.$$

其精确解是 $y = 2/(1+x^2)$.

解 ☞ 这个解开始时相对陡地朝下转，然后逐渐变成水平. 所以我们预料开始需要小的步长，以后可以逐渐放宽. 观察进行到 $x=27$ 所预期的这些步长很有趣.

x	0.15	1	2	3	4	9	12	17	27
h	0.07	0.05	0.1	0.2	0.3	0.9	1.4	2.7	4.3

19.49 什么是可变阶方法？

解 ☞ 用于对微分方程进行积分的公式的可变阶，是试图用最少的计算达到给定的精度的另一种方法. 开始用一个低阶公式作为自启动，这需用小的步长使它精确，在进行计算时两者进行调整. 想法是对当前计算的这一步找出最优阶和步长. 为了做到这一点，已有许多专门的程序，它们都有些复杂. 但是其构成策略的基础类似于题 19.44 到题 19.46.

刚性方程

19.50 什么是刚性方程？

解 ☞ 这个术语一般和方程组联系在一起. 但原则上能比较简单地说明. 取方程

$$y' = -100y + 99e^{-x},$$

它有解

$$y = e^{-x} - e^{-100x},$$

满足初始条件 $y(0)=0$. 这个解的两项都趋于零，但是第二项比第一项衰减快得多. 在 $x=0.1$ 处第二项四位小数都已为零. 与第一项相比它的确是瞬变项. 而第一项几乎就可被称为"稳态". 方程组中不同分量以十分不同的时间尺度作运算，这种方程组就称为刚性方程组，它们与正常的数值解大相庭径.

19.51 考虑到前面的瞬变项迅速衰减，我们可以预期用步长 $h=0.1$ 来生成留下的项 e^{-x}的值. 经典的 Runge-Kutta 方法实际产生了什么？

解 ☞ 很像题19.42.我们有 $f_y=-100$,把它与稳定性准则中的 A 相结合,就成为

$$-2.78\leqslant Ah=-100h.$$

这就启发我们保持步长 h 比0.0278小.这可能有些吃惊,因为它似乎在暗示瞬变项(在 $x=0.1$ 后其大小可不计.)仍能以一种重要的、隐秘的方法影响计算.将这个理论进行试验,用 $h=0.03$ 进行运算.预言的"淹没"并没有出现,y 的值很快下降到 -10^{14},但是用 $h=0.025$ 就导致成功的一轮运算.在 $x=3$ 处得出0.04980,这恰好在第五位处高出一个单位.

19.52 建立 Gear 公式

$$\nabla y_{n+1}+\frac{1}{2}\nabla^2 y_{n+1}+\frac{1}{3}\nabla^3 y_{n+1}=hy'_{n+1},$$

这里∇是向后差分算子.证明它等价于

$$y_{n+1}=\frac{18}{11}y_n-\frac{9}{11}y_{n-1}+\frac{2}{11}y_{n-2}+\frac{6h}{11}y'_{n+1},$$

这里

$$y'_{n+1}=f(x_{n+1},y_{n+1}).$$

解 ☞ 以 Newton 向后公式开始

$$p_k=y_{n+1}+k\nabla y_{n+1}+\frac{k(k+1)}{2}\nabla^2 y_{n+1}+\frac{k(k+1)(k+2)}{6}\nabla^3 y_{n+1}$$

(见题7.9),其中 $x-x_{n+1}=kh$,p_k 是 k 的三次多项式,它与 y 在 $k=0,-1,-2,-3$ 相配置,我们求导并令 $k=0$

$$\left.\frac{\mathrm{d}p}{\mathrm{d}x}\right|_{k=0}=\left.\frac{\mathrm{d}p}{\mathrm{d}k}\frac{1}{h}\right|_{k=0}=\frac{1}{h}\left(\nabla y_{n+1}+\frac{1}{2}\nabla^2 y_{n+1}+\frac{1}{3}\nabla^3 y_{n+1}\right).$$

把它取作 y'_{n+1} 的近似,我们就得到了第一个 Gear 公式.第二个公式容易通过把向后差分算子用 y_i 表示而得到.第二个公式也能用待定系数法得到.即要求它对直到三次的多项式精确成立.相应的高阶公式可通过推广而得到.例如把 Newton 公式延伸到 $k=-4$,引入四阶导数项即在上式左边加入 $\frac{1}{4}\nabla^4 y_{n+1}$.

19.53 为什么 Gear 公式能用来解刚性问题?

解 ☞ 对于比我们的其他公式的 h 值大得多时它们证明是稳定的.再用题19.50中的方程为例,我们已发现对于 $h=0.03$ Runge-Kutta 方法不稳定,可是 Gear 公式现化为

$$y_{n+1}=\frac{18y_n-9y_{n-1}+2y_{n-2}+594he^{-(x_n+h)}}{11+600h}$$

(用方程中的 y' 代入,然后解得 y_{n+1}),取 $h=0.1$(用三个精确的起初值)可得到

x	2	4	6
y	0.135336	0.018316	0.002479

其中第一个在最后一位高了一个单位.至于 $h=0.5$ 可以认为是最合适的结果

x	2	4	6
y	0.1350	0.01833	0.002480

较大的 h 产生较大的截断误差,但不会产生稳定性的麻烦.

19.54 Gear 公式一般对 y_{n+1} 是非线性的,建立 Newton 迭代并用它求解这个未知量.

解 ☞ 上例中 $f(x,y)$ 关于 y 是线性的,于是可以直接解出 y_{n+1}.但是一般地我们必须考虑当

$$F(y)=y-\frac{6h}{11}f(x_{n+1},y)-S=0$$

时的 Gear 公式,其中已把 y_{n+1} 简写为 y,S 表示不包括 y_{n+1} 的三项之和.于是 Newton 迭代为

$$y^{(k+1)} = y^{(k)} - \frac{F(y^{(k)})}{F'(y^{(k)})},$$

其中　$F'(y) = 1 - \frac{6h}{11} f_y(x_{n+1}, y)$.

补 充 题

19.55　通过考虑方程 $y' = x^2 - y^2$ 的方向场，导出其解的性态.解在什么地方取极大和极小？在什么地方它们的曲率为 0？证明对大的正数 x，必有 $y(x) < x$.

19.56　对上一题中的方程试用图解法估计经过 $(-1,1)$ 的解 y 在 $x=0$ 处等于什么？

19.57　通过考虑方程 $y' = -2xy$ 的方向场，导出其解的性态.

19.58　对 $y' = -xy^2$, $y(0) = 2$，用简单 Euler 方法，取 $h = 0.5, 0.2, 0.1, 0.01$ 一直计算到 $x = 1$.这些解是否向精确解 $y(1) = 1$ 收敛.

19.59　对 $y' = -xy^2$, $y(0) = 2$ 用"中点公式" $y_{k+1} \approx y_{k-1} + 2hf(x_k, y_k)$ 取 $h = 0.1$ 并证明 $y(1) \approx 0.9962$.

19.60　对 $y' = -xy$, $y(0) = 2$ 用修正 Euler 公式，并对以上三个问题中所得预测结果进行比较.对相同的 h，这些很简单的方法中哪一个进行得最好？你能解释为什么吗？

19.61　对于 $y' = -xy^2$, $y(0) = 2$，取 $h = 0.2$ 用局部 Taylor 级数方法求解.把你的结果和已解得问题的结果进行比较.

19.62　把 Runge-Kutta 方法用于上述问题，并比较你的结果.

19.63　证明题 19.9 中的第一个公式.

19.64　对 $y' = xy^{1/3}$, $y(1) = 1$，取 $h = 0.1$ 用 Milne 预测-校正方法求解，把所得结果和已解决问题中的结果进行比较.

19.65　对上述问题用 Adams 预测-校正方法求解，再比较结果.

19.66　对题 19.64，采用二个或三个其他预测-校正组合.其结果是否有本质的差别？

19.67　对 $y' = x^2 - y^2$, $y(-1) = 1$ 用各种方法求解.$y(0)$ 等于什么？这和你在题 19.56 中所作的估计有多接近？

19.68　用各种方法求解 $y' = -2xy$, $y(0) = 1$，与精确解 $y = e^{-x^2}$ 进行比较结果怎样？

19.69　证明把 Milne 方法用于 $y' = y$, $y(0) = 1$ 取 $h = 0.3$ 并取四位小数，导出以下相对误差：

x	1.5	3.0	4.5	6.0
相对误差	0.00016	0.00013	0.00019	0.00026

这意味着计算几乎总是产生四位有效数字.

19.70　证明把 Milne 方法用于 $y' = -y$, $y(0) = 1$，取 $h = 0.3$ 并取五位小数，导出以下相对误差：

x	1.5	3.0	4.5	6.0
相对误差	0	−0.0006	0.0027	−0.0248

虽然几乎产生四位准确数字，但是相对误差已开始其不断增长的振荡.

19.71　证明中点方法

$$Y_{k+1} = Y_{k-1} + 2hf(x_k, Y_k)$$

是不稳定的.

证明这个公式比 Euler 方法有较低的截断误差，其精确解满足

$$y_{k+1} = y_{k-1} + 2hf(x_k, y_k) + \frac{1}{3}h^3 y^{(3)}(\xi).$$

对于特殊情形 $f(x,y) = Ay$，证明

$$d_{k+1} = d_{k-1} + 2hAd_k.$$

为了再次集中单个误差 d_0 的长期影响，不计及截断误差项.

为了解这个方程，先证明 $r^2-2hAr-1=0$ 的两个根是

$$r = hA \pm \sqrt{h^2A^2+1} - hA = 1 + 0(h^2).$$

对小的 hA，它们接近 e^{hA} 和 e^{-hA}. 因此差分方程的解是

$$d_k = c_1(1+Ah)^k + c_2(-1)^k(1-Ah)^k \approx c_1e^{Ahk} + c_2(-1)^ke^{-Ahk}.$$

令 $k=0$ 得到 $d_0=c_1+c_2$，除以 y_k，相对误差为

$$r_k \approx c_1 + (d_0-c_1)(-1)^ke^{-2Ahk}.$$

证明对正的 A，它将保持有界，但对负的 A，当 k 增加时，它将成为无界，因此这个方法在这种情形是不稳定的.

19.72 表 19.6 中的结果是对方程 $y'=-xy^2, y(0)=2$ 应用中点方法所得到的. 步长 $h=0.1$，仅对 $x=0.5(0.5)5$ 的值进行打印. 这个方程不是线性的，但计算每一个值的相对误差发现通过前面线性问题的分析预报振荡迅速地增加.

表 19.6

x_k	计算 y_k	正确 y_k	x_k	计算 y_k	正确 y_k
0.5	1.5958	1.6000	3.0	0.1799	0.2000
1.0	0.9962	1.0000	3.5	0.1850	0.1509
1.5	0.6167	0.6154	4.0	0.0566	0.1176
2.0	0.3950	0.4000	4.5	0.1689	0.0941
2.5	0.2865	0.2759	5.0	0.0713	0.0769

19.73 对题 19.27 中列出的其他的校正公式分析其相对误差.

19.74 证明公式 $y_{k+1} \approx y_k + \frac{1}{2}h(y'_{k+1}+y'_k) + \frac{1}{12}h^2(y''_{k+1}+y''_k)$

有截断误差 $h^5y^{(5)}(\xi)/720$，而类似的预测公式

$$y_{k+1} \approx y_k - \frac{1}{2}h(-y'_k + 3y'_{k-1}) + \frac{1}{12}h^2(17y''_k + 7y''_{k-1})$$

有截断误差 $31h^5y^{(5)}(\xi)/6!$. 这些公式用第二个导数的值来降低截断误差.

19.75 对 $y'=-xy^2, y(0)=2$，取 $h=0.2$ 应用上题中的公式，需要用一个额外的起始值，它可以取为同一方程以前的解，譬如说由 Taylor 级数.

19.76 给定 $y'=\sqrt{1-y^2}, y(0)=0$，作为试验，用我们的近似方法中的任一种方法计算 $y(\pi/2)$.

19.77 给定 $y'=x-y, y(0)=2$，用我们的近似方法中的任一种求 $y(2)$.

19.78 用任一种我们的近似方法求解 $y'=\dfrac{y(1-x^2y^4)}{x(1+x^2y^4)}, y(1)=1$，直到 $x=2$.

19.79 用任一种我们的近似方法求解 $y'=-\dfrac{2xy+e^y}{x^2+xe^y}, y(1)=0$，直到 $x=2$.

19.80 用任一种我们的近似方法求解 $y'=\dfrac{2x+y}{2y-x}, y(1)=0$，直到 $x=2$.

19.81 一物体降落至地球的过程中，根据 Newton 理论，仅考虑地球的重力吸引，按照方程（同时参看题 20.16）

$$\frac{dy}{dt} = -\sqrt{2gR^2}\sqrt{\frac{H-y}{Hy}}.$$

这里 y = 到地球中心的距离，$g=32$，$R=4000(5280)$，H = 到地球中心的初始距离. 可以证明这个方程的精确解是

$$t = \frac{H^{3/2}}{8y}\left[\sqrt{\frac{y}{H}-\left(\frac{y}{H}\right)^2} + \frac{1}{2}\arccos\left(\frac{2y}{H}-1\right)\right],$$

初始速度为 0，把一种我们的近似方法用于这个微分方程及初始条件 $y(0)=H=237,000(5280)$. 在什么时刻你求得 $y=R$？如果月球在其路程中停止并且地球保持稳定，这个结果可被解释为是月球降落到地球所要的时间.

19.82 质量为 m 的雨滴在下落时间 t 后有速度 v. 设运动方程是

$$\frac{dv}{dt} = 32 - \frac{cv^2}{m}.$$

其中 c 是空气阻力的一种测量,于是能证明速度趋于一个有限值.直接应用一种我们的近似方法到这个微分方程中来证实这一结论(对于情形 $c/m=2$)可用任一初始速度.

19.83　子弹克服空气阻力 cv^2 向上发射,假设运动方程是

$$\frac{dv}{dt}=-32-\frac{cv^2}{m}.$$

如果 $c/m=2$, $v(0)=1$,用一种我们的方法来求子弹达到最高点所需要的时间.

19.84　长度为 L 的绳的一端支承在一条直线上,连接另一端的重物的路程由下式决定(见图 19.7)

$$y'=-\frac{y}{\sqrt{L^2-y^2}},$$

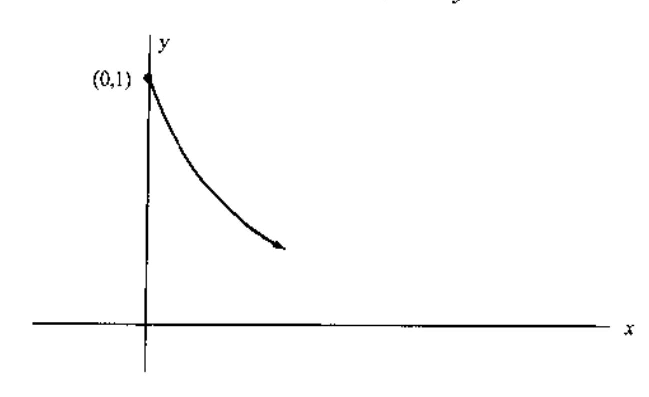


图 19.7

能求出精确解.但是用一种我们的近似方法来计算重物的路程,从 $(0,L)$ 出发,取 $L=1$.

第二十章　高阶微分问题

基本问题

本章考虑的基本问题是对形如

$$y_i' = f_i(x, y_1, \cdots, y_n) \qquad i = 1, \cdots, n$$

的一阶微分方程组和给定的初始条件 $y_i(x_0) = a_i$，求 n 个函数 $y_i(x)$. 它出现在各种应用中. 这是第十九章处理的初值问题的直接推广. 把它写成向量形式使其特别简单.

$$Y'(x) = F(x, Y), \qquad Y(x_0) = A,$$

其中 Y, F 和 A 分别有分量 y_i, f_i 和 a_i.

高阶微分方程可以用这一阶微分方程组来代替，而且有标准的处理方法. 作为最简单的例子，二阶方程

$$y'' = f(x, y, y')$$

化为二个函数 y 和 p 的方程组

$$y' = p, \qquad p' = f(x, y, p).$$

相伴的初始条件 $y(x_0) = a, y'(x_0) = b$ 由 $y(x_0) = a$ 和 $p(x_0) = b$ 所代替. 这样就得到了上述基本问题. 对于三阶方程定义 $y' = p$ 和 $y'' = q$ 很快就导出三个方程的一阶方程组，如此等等. 高阶方程组如上述那样处理每一项. 因此把任何一个高阶问题化为一阶方程组的这种选取是可用的.

求解方法

容易把上一章的方法推广到一阶方程组，Taylor 级数的应用很直接，因此常适用于此，Runge-Kutta 方法也可应用，方程组中每一个方程几乎完全和在第十九章中那样处理. 预测-校正方法同样正确，这些推广的例子将在题解中给出.

题　解

20.1　通过解方程组

$$x' = -x - y,$$
$$y' = x - y.$$

来说明解联立方程组的 Taylor 级数方法，两个函数 $x(t)$ 和 $y(t)$ 满足初始条件 $x(0) = 1, y(0) = 0$.

解　直接代入两级数

$$x(t) = x(0) + tx'(0) + \frac{1}{2}t^2x''(0) + \cdots,$$

$$y(t) = y(0) + ty'(0) + \frac{1}{2}t^2y''(0) + \cdots.$$

由所给的方程组得到需要的部分，首先 $x'(0) = -1$ 及 $y'(0) = 1$，然后从 $x'' = -x' - y'$ 和 $y'' = x' - y'$ 得到 $x''(0) = 0, y''(0) = -2$. 同样的方法可以得到高阶导数. 这两个级数开始如下，

$$x(t) = 1 - t + \frac{1}{3}t^3 - \frac{1}{6}t^4 + \cdots,$$

$$y(t) = t - t^2 + \frac{1}{3}t^3 + \cdots.$$

所给方程组不但是线性的而且有常系数，把它写成如下形式

$$X'(t) = AX(t),$$

其中
$$X' = \begin{pmatrix} x \\ y \end{pmatrix}, \qquad A = \begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix}.$$

用
$$X = e^{\lambda t}\begin{pmatrix} a \\ b \end{pmatrix}$$

代入试验就能求得精确解,代入方程就得到了矩阵 A 的特征值问题.对上面的 A 有

$$(-1-\lambda)a - b = 0,$$
$$a + (-1-\lambda)b = 0.$$

得到 $\lambda = -1 \pm i$,稍作计算就得到

$$x(t) = e^{-t}\cos t, \qquad y(t) = e^{-t}\sin t.$$

上面开始的 Taylor 级数当然就是这些函数的级数.

所述的过程容易推广到较大的方程组.

20.2 用经典的四阶集对二个联立一阶方程组写出 Runge-Kutta 公式.

解 ☞ 设所给方程是

$$y' = f_1(x,y,p), \qquad p' = f_2(x,y,p).$$

初始条件是 $y(x_0) = y_0, p(x_0) = p_0$.可以证明以下公式

$$k_1 = hf_1(x_n, y_n, p_n), \qquad k_3 = hf_1\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_2, p_n + \frac{1}{2}l_2\right),$$

$$l_1 = hf_2(x_n, y_n, p_n), \qquad l_3 = hf_2\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_2, p_n + \frac{1}{2}l_2\right),$$

$$k_2 = hf_1\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1, p_n - \frac{1}{2}l_1\right), k_4 = hf_1(x_n + h, y_n + k_3, p_n + l_3),$$

$$l_2 = hf_2\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1, p_n - \frac{1}{2}l_1\right), l_4 = hf_2(x_n + h, y_n + k_3, p_n - l_3),$$

$$y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

$$p_{n+1} = p_n + \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4).$$

等同于关于两个函数直到四阶项的 Taylor 级数,其细节与单个方程完全相同,因此略去.对多于两个,譬如 n 个联立方程组,Runge-Kutta 方法的推广和上面是平行的.用 n 组公式代替两组公式.这种公式的应用之例见题 20.7.

20.3 对上题中的联立方程组写出 Adams 型的预测-校正公式.

解 ☞ 假设已有每一函数的四个初始值,譬如 y_0, y_1, y_2, y_3 和 p_0, p_1, p_2, p_3,于是可用预测公式

$$y_{k+1} \approx y_k + \frac{1}{24}h(55y'_k - 59y'_{k-1} + 37y'_{k-2} - 9y'_{k-3}),$$

$$p_{k+1} \approx p_k + \frac{1}{24}h(55p'_k - 59p'_{k-1} + 37p'_{k-2} - 9p'_{k-3}),$$

以及
$$y'_k = f_1(x_k, y_k, p_k), \qquad p'_k = f_2(x_k, y_k, p_k).$$

这些结果可用来启动校正公式

$$y_{k+1} \approx y_k + \frac{1}{24}h(9y'_{k+1} + 19y'_k - 5y'_{k-1} + y'_{k-2}),$$

$$p_{k+1} \approx p_k + \frac{1}{24}h(9p'_{k+1} + 19p'_k - 5p'_{k-1} + p'_{k-2}).$$

于是重复进行到相邻结果和指定的允许偏差相一致.此过程几乎和单个方程的过程没有不同.推广到更多的方程或其他的预测-校正组合是类似的.

高阶方程如同方程组

20.4 证明二阶微分方程能用两个一阶方程的方程组代替.

证 ☞ 设二阶方程是 $y'' = f(x, y, y')$.引入 $p = y'$,我们立即有 $y' = p, p' = f(x, y, p)$.作为这一标准方法的结果.只要需要,一个二阶方程就可以用方程组的方法来处理.

20.5 证明一般的 n 阶方程

$$y^{(n)} = f(x, y, y', y^{(2)}, \cdots, y^{(n-1)})$$

也可以用一阶方程组代替.

证 ☞ 为方便计,把 $y(x)$ 称为 $y_1(x)$,并引入附加的函数 $y_2(x), \cdots, y_n(x)$

$$y_1' = y_2, \quad y_2' = y_3, \quad \cdots, \quad y_{n-1}' = y_n.$$

那么原 n 阶方程成为

$$y_n' = f(x, y_1, y_2, \cdots, y_n).$$

这 n 个方程是一阶的,并且可用方程组的方法求解.

20.6 用等价的一阶方程组代替三维质点运动方程

$$x'' = f_1(t, x, y, z, x', y', z'), \qquad y'' = f_2(t, x, y, z, x', y', z'),$$
$$z'' = f_3(t, x, y, z, x', y', z').$$

解 ☞ 设 $x' = u, y' = v, z' = w$ 是速度分量,于是

$$u' = f_1(t, x, y, z, u, v, w), \qquad v' = f_2(t, x, y, z, u, v, w),$$
$$w' = f_3(t, x, y, z, u, v, w).$$

这六个方程就是所要求的一阶方程组.其他高阶方程可以用同样方法处理.

20.7 计算 Van der Pol 方程

$$y'' - (0.1)(1 - y^2)y' + y = 0$$

初始条件是 $y(0) = 1, y'(0) = 0$ 的解直到 $y(t)$ 的第三位小数.对两个一阶方程使用 Runge-Kutta 公式.

解 ☞ 等价的一阶方程组是

$$y' = p = f_1(t, y, p),$$
$$p' = -y + (0.1)(1 - y^2)p = f_2(t, y, p).$$

这个方程组的 Runge-Kutta 公式是

$$k_1 = hp_n, \qquad l_1 = h[-y_n + (0.1)(1 - y_n^2)p_n],$$
$$k_2 = h\left(p_n + \frac{1}{2}l_1\right), \quad l_2 = h\left\{-\left(y_n + \frac{1}{2}k_1\right) + (0.1)\left[1 - \left(y_n + \frac{1}{2}k_1\right)^2\right]\left(p_n + \frac{1}{2}l_1\right)\right\},$$
$$k_3 = h\left(p_n + \frac{1}{2}l_2\right), \quad l_3 = h\left\{-\left(y_n + \frac{1}{2}k_2\right) + (0.1)\left[1 - \left(y_n + \frac{1}{2}k_2\right)^2\right]\left(p_n + \frac{1}{2}l_2\right)\right\},$$
$$k_4 = h(p_n + l_3), \qquad l_4 = h\{-(y_n + k_3) + (0.1)[1 - (y_n + k_3)^2](p_n + l_3)\},$$

以及 $\quad y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \quad p_{n+1} = p_n + \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4)$

取 $h = 0.2$ 计算得到以下三位小数的结果

$$k_1 = (0.2)(0) = 0, \qquad l_1 = (0.2)[-1 + (0.1)(1-1)(0)] = -0.2,$$
$$k_2 = (0.2)(-0.1) = -0.02, \qquad l_2 = (0.2)[-1 + (0.1)(1-1)(-0.1)] = -0.2,$$
$$k_3 \approx (0.2)(-0.1) = -0.02, \qquad l_3 = (0.2)[-0.99 + (0.1)(0.02)(-0.1)] \approx -0.198,$$

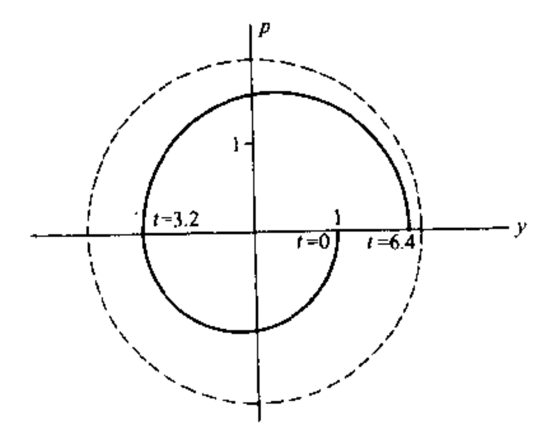


图 20.1

$k_4 \approx (0.2)(-0.198) \approx -0.04,\ \ l_4 = (0.2)[-(0.98)+(0.1)(0.04)(-0.198)] \approx -0.196.$
把这些值代入得到

$$y_1 \approx 1 + \frac{1}{6}(-0.04 - 0.04 - 0.04) = 0.98,$$

$$p_1 \approx 0 + \frac{1}{6}(-0.2 - 0.4 - 0.396 - 0.196) \approx -0.199.$$

$n=1$ 时第二步继续用这种方法计算. 在图 20.1 中的图像说明直到 $t=6.4$ 的结果, 这时曲线再次向下穿过坐标轴, 这里 y 和 p 是坐标. 这种"相图"常用于振荡系统的研究. 这里振荡(实线表示)变强, 而且当 x 趋于无穷时趋于周期振荡(虚线表示), 这一点在非线性振荡理论中得到证明.

用级数求解高阶方程

20.8 在 $x=0$ 的邻域内求线性方程 $y''+(1+x^2)y=e^x$ 的级数解:

解 ☞ 设级数为 $y(x)=\sum_{i=0}^{\infty} a_i x^i$ 代入可得

$$\sum_{i=2}^{\infty} a_i i(i-1)x^{i-2} + (1+x^2)\sum_{i=0}^{\infty} a_i x^i = \sum_{i=0}^{\infty} \frac{x^i}{i!}.$$

它经过改变下标可转变为

$$(a_0+2a_2)+(a_1+6a_3)x+\sum_{k=2}^{\infty}[(k+2)(k+1)a_{k+2}+a_k+a_{k-2}]x^k = \sum_{k=0}^{\infty}\frac{x^k}{k!}.$$

比较 x 的同次幂的系数得 $a_2=(1-a_0)/2,\ a_3=(1-a_1)/6$, 于是递推可得

$$(k+2)(k+1)a_{k+2} = -a_k - a_{k-2} - \frac{1}{k!}.$$

逐次可得 $a_4=-a_0/24,\ a_5=-a_1/24,\ a_6=(13a_0-11)/720$ 等等. 数 a_0 和 a_1 将由初始条件决定. 因为我们的微分方程各部分是解析函数, 在自变量 x 附近可建立相似的级数, 这些级数可能已足以计算整个所需区间的解, 如果不能, 就用它为其他方法来生成启动值.

20.9 在 $x=0$ 的邻域内求非线性方程 $y''=1+y'^2,\ y(0)=y'(0)=0$ 的级数解.

解 ☞ 能够用前题中的方法, 但将再次说明要直接计算高阶函数. 容易计算

$$y^{(3)}=2yy',\qquad y^{(4)}=2y(1+y^2)+2(y')^2,\qquad y^{(5)}=10y^2y'+6y',$$
$$y^{(6)}=20(y')^2+(1+y^2)(10y^2+6)$$

等等. 根据所给的初始条件, 它们在零点的值除了 $y^{(6)}$ 以外都是零, 因此由 Taylor 定理得到

$$y=\frac{1}{2}x^2+\frac{1}{120}x^6+\cdots.$$

20.10 对刚性方程组

$$y'=p,$$
$$p'=-100y-101p$$

和初始条件 $y(0)=1,\ p(0)=-1$, 应用题 19.52 中的 Gear 方法. 这个方程组等价于二阶方程

$$y''+101y'+100y=0,$$
$$y(0)=1,\ y'(0)=1.$$

其精确解是 $y(x)=e^{-x}$.

解 ☞ Runge-Kutta 方法能处理这一方程. 但是为了保证是一种稳定的计算, 经典的四阶集合需用小于 0.0278 的步长, 对 y 和 p 写出 Gear 公式

$$y_{n+1}=\frac{1}{11}(18y_n-9y_{n-1}+2y_{n-2})+\frac{6h}{11}p_{n+1},$$

$$p_{n+1}=\frac{1}{11}(18p_n-9p_{n-1}+2p_{n-2})+\frac{6h}{11}(-100y_{n+1}-101p_{n+1}).$$

这可写成关于 y_{n+1} 和 p_{n+1} 的线性方程组

$$y_{n+1}-\frac{6h}{11}p_{n+1}=\frac{1}{11}(18y_n-9y_{n-1}-2y_{n-2}),$$

$$\frac{600h}{11}y_{n+1}+\left(1+\frac{606h}{11}\right)p_{n+1}=\frac{1}{11}(18p_n-9p_{n-1}+2p_{n-2}).$$

因为方程组是线性的，所以不需用 Newton 迭代求解．下面给出了选取两种步长 h 的结果．这两种步长都比 Runge-Kutta 方法所需用的步长大得多，为了进行比较还列出了正确值．

x	$y=e^{-x}$	$h=0.1$	$h=0.2$
2	0.1353	0.1354	0.1359
4	0.01832	0.01833	0.0185
6	0.002479	0.002483	0.00251
8	0.0003355	0.0003362	0.000342
10	0.0000454	0.0000455	0.000465

20.11　一条在外面田野上的狗，看见它的主人沿道路散步就奔向他．假设狗始终直接瞄准它的主人，而且路是直的，确定狗的路程的方程（见图 20.2）是

$$xy''=c\sqrt{1+(y')^2},$$

其中 c 是人速与狗速之比．根据著名的攻击线（line of attack）得出精确解

$$y=\frac{1}{2}\left(\frac{x^{1+c}}{1+c}-\frac{x^{1-c}}{1-c}\right)+\frac{c}{1-c^2},$$

其中 c 小于 1．当 x 趋于零，狗在 $y=c/(1-c^2)$ 处追上它的主人，对于 $c=\frac{1}{2}$ 的情形，用近似方法来解这个问题．追赶在 $y=\frac{2}{3}$ 处结束．

解　☞ 这个二阶方程首先用以下方程组

$$y'=p,$$

$$p'=\frac{c\sqrt{1+p^2}}{x}$$

表 20.1

x	y
0.1	0.3608
0.01	0.5669
0.001	0.6350
0.0001	0.6567
0.00001	0.6636
0.0000006	0.6659
-0.0000003	0.6668

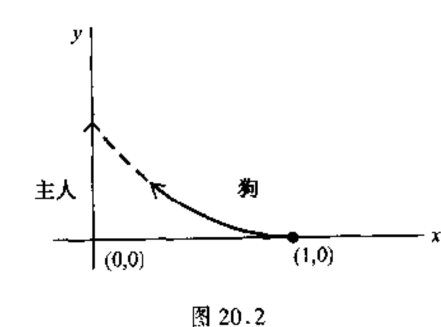


图 20.2

来代替．初始条件是 $y(1)=0$，$p(1)=0$．能再一次用题 20.2 中的 Runge-Kutta 公式，这时取负的 h 值．这里惟一的困难是当 x 接近零时斜率 p 变得很大，步长 h 递减的自适应方法看来是必要的．我们着力于开始的策略，先取 $h=-0.1$ 直到 $x=0.1$，再取 $h=-0.01$ 直到 $x=0.01$，等等，结果列于表 20.1 中．最后两个 x 的值看来包括舍入误差，p 的值没有列出，但是在大小上升至接近 1000．

20.12　方程

$$r''=\frac{9}{r^3}-\frac{2}{r^2},\qquad \theta'=\frac{3}{r^2}.$$

其中一撇，二撇都表示对 t 求导数，是在适当选取某些物理常数后描述质点在平方反比重力场中的 Newton 轨道．如果在 r 的最小值位置 $t=0$（图 20.3），而且

$$r(0)=3,\qquad \theta(0)=0,\qquad r'(0)=0.$$

于是这轨道证明是椭圆 $r=9/(2+\cos\theta)$．用一种我们的近似方法并与精确结果进行

比较.

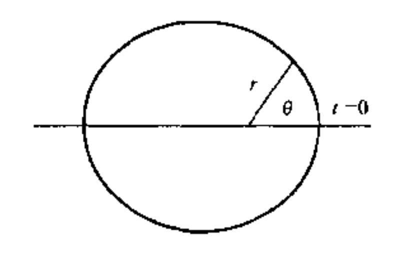


图 20.3

解 ☞ 应用十分直接.用熟悉的化为一阶方程组的方法先把它化为

$$r' = p, \qquad p' = \frac{9}{r^3} - \frac{2}{r^2}, \qquad \theta' = \frac{3}{r^2},$$

再加上 Runge-Kutta 方法的三组程序,仍按照题 20.2 中的模型,积分继续到角度 θ 超过 2π.在表 20.2 中提供了输出值有选择的摘录(用步长 $h=0.1$),它明显地具所要轨道的性质.进一步检查,理论给出了周期 $=12\pi\sqrt{3}$ 或者大约是 65.3.而这一点符合得很好.

表 20.2

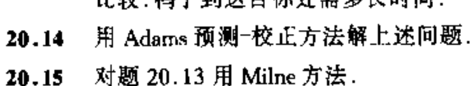
t	r	θ	p
0	3.00	0.00	0.00
6	4.37	1.51	0.33
7	4.71	1.66	0.33
32	9.00	3.12	0.01
33	9.00	3.15	−0.004
59	4.47	4.73	−0.33
65	3.00	6.18	−0.03
66	3.03	6.52	0.08

补　充　题

20.13 方程

$$x'(t) = -\frac{2x}{\sqrt{x^2+y^2}}, \qquad y'(t) = 1 - \frac{2y}{\sqrt{x^2+y^2}}.$$

描述了鸭子始终瞄准目标位置而企图游过一条河的路程,水流速度为 1,鸭子的速度为 2,鸭子在 S 处出发,因此 $x(0)=1$, $y(0)=0$(见图 20.4),对两个齐次方程用 Runge-Kutta 公式计算鸭子的路程.与精确解 $y=\frac{1}{2}(x^{1/2}-x^{2/1})$ 进行比较.鸭子到达目标处需多长时间?



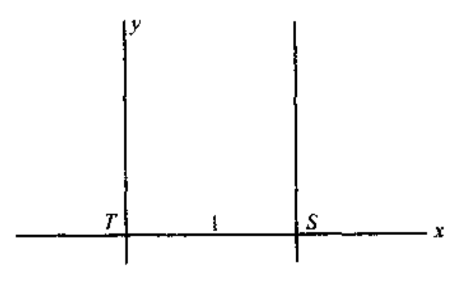


图 20.4

20.14 用 Adams 预测-校正方法解上述问题.

20.15 对题 20.13 用 Milne 方法.

20.16 一个物体下降至地球的经典的平方反比定律是

$$y''(t) = -\frac{gR^2}{y^2},$$

其中 g 是常数,R 是地球半径.它有著名的而且多少让人吃惊的解

$$t = \frac{H^{3/2}}{8y}\left[\sqrt{\frac{y}{H}-\left(\frac{y}{H}\right)^2} + \frac{1}{2}\arccos\left(\frac{2y}{H}-1\right)\right],$$

其中 H 是初始高度,初始速度为零.引入等价方程组

$$y' = p, \qquad p' = -\frac{gR^2}{y^2}.$$

用 Runge-Kutta 方法计算速度 $p(t)$ 和位移 $y(t)$.何时自由落体到达地球表面?并与精确解比较.(若用英里和秒作单位,则 $g=\frac{32}{5280}$,$R=4000$,并取 $H=200,000$,即月球到地球的距离.这个问题说明了某些计算空间轨道的困难.)

20.17 对题 20.16 用 Adams 方法.

20.18 证明 $yy''+3(y')^2=0$, $y(0)=1$, $y'(0)=\frac{1}{4}$ 的解可表为

$$y(x) = 1 + \frac{x}{4} - \frac{3x^2}{32} + \frac{7x^3}{128} - \frac{77x^4}{2048} + \cdots.$$

20.19　证明 $x^2y'' - 2x^2y' + \left(\frac{1}{4} + x^2\right)y = 0$ 有以下形式的解

$$y(x) = \sqrt{x}(a_0 + a_1x + a_2x^2 + \cdots).$$

如果要求 $\lim_{x\to 0}\frac{y(x)}{\sqrt{x}} = 1$，试确定其系数.

20.20　把 Runge-Kutta 公式用于

$$y' = -12y + 9z, \qquad z' = 11y - 10z.$$

使用 $y(1)\approx 9e^{-1}$，$z(1)\approx 11e^{-1}$ 作为初始条件，

这个方程组有精确解　$y = 9e^{-x} + 5e^{-21x}$，　$z = 11e^{-x} - 5e^{-21x}$.

取 $h = 0.2$ 计算到三位或四位小数，而且至少算到 $x = 3$. 注意如果 $\frac{11y}{9z}$ 接近 1，那么计算便开始严重振荡. 要解释这一点可通过对 e^{-21x} 的四次 Taylor 近似（这是 Runge-Kutta 方法必须使用的）和精确的指数函数相比较.

第二十一章　最小二乘多项式逼近

最小二乘原则

选择一个多项式 $p(x)$用一种使平方误差(在某种意义下)为极小的方式来逼近一个已知函数,其基本思想首先为 Gauss 所提出.有若干种版本,取决于所涉及的自变量的集合及所用的误差度量.

首先,当数据为离散时,我们可以对给定的数据 x_i, y_i 及 $m<N$ 将和

$$S = \sum_{i=0}^{N}(y_i - a_0 - a_1x_i - \cdots - a_mx_i^m)^2$$

极小化.条件 $m<N$ 使多项式

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_mx^m$$

未必能在所有 N 个数据点上都能匹配.故 S 很可能不会变成零.Gauss 思想是尽我们所能使 S 变小.于是微积分学中的标准技法便引出了决定 a_j 的法方程组.这些方程是

$$\begin{aligned}
&s_0a_0 + s_1a_1 + \cdots + s_ma_m = t_0,\\
&s_1a_0 + s_2a_1 + \cdots + s_{m+1}a_m = t_1,\\
&\cdots\\
&s_ma_0 + s_{m+1}a_1 + \cdots + s_{2m}a_m = t_m.
\end{aligned}$$

其中 $s_k = \sum_{i=0}^{N}x_i^k, t_k = \sum_{i=0}^{N}y_ix_i^k$.这个线性方程组的确能惟一地决定 a_i 而且所得到的a_i 确实产生 S 的最小可能值.对于线性多项式来说

$$p(x) = Mx + B,$$

法方程容易解出并且得到

$$M = \frac{s_0t_1 - s_1t_0}{s_0s_2 - s_1^2}, \qquad B = \frac{s_2t_0 - s_1t_1}{s_0s_2 - s_1^2}.$$

为了对现存的不同的最小二乘方法提供一个统一的处理,包括刚才描述过的第一种方法,考虑在向量空间中极小化的一般问题.从代数观点,用正交投影思想易得其解.自然地这个一般问题再现我们的 $p(x)$及法方程.当我们进行解其他变种的最小二乘原则时将会重新作出解释.对所着手的特殊情况在大多数场合还会重复提供同样的论点.

除次数非常低的多项式之外,上面的法方程组被证明为病态(坏条件)的.这说明了,虽然它惟一地定义系数 a_j,但在实践中可以证明这些 a_j 不可能被解出.(将在第 26 章中提出的)解线性方程的标准方法或者根本不可能产生解,或者数据误差被恶性放大.为解决问题,正交多项式被引进.(这相当于为抽象的向量空间选择一个正交基.)对于离散数据的情况,它们是次数为 $m=0,1,2,\cdots$的多项式 $P_{m,N}(t)$,具有性质

$$\sum_{t=0}^{N}P_{m,N}(t)P_{n,N}(t) = 0,$$

这就是正交性.显式的表达式

$$P_{m,N}(t) = \sum_{i=0}^{m}(-1)^i\binom{m}{i}\binom{m+i}{i}\frac{t^{(i)}}{N^{(i)}}$$

将被得到,在这个式子中二项式系数和阶乘多项式显得重要.

我们的最小二乘多项式的另一种形式现在变得方便了,即

$$p(t) = \sum_{k=0}^{m}a_kP_{k,N}(t)$$

含有新系数 a_k. 决定这些 a_k 的方程组证明为特别容易求解. 事实上

$$a_k = \frac{\sum_{t=0}^{N} y_t P_{k,N}(t)}{\sum_{t=0}^{N} P_{k,N}^2(t)},$$

这些 a_k 的确使误差和 S 极小化, 极小值为

$$S_{\min} = \sum_{t=0}^{N} y_t^2 - \sum_{k=0}^{m} W_k a_k^2,$$

其中 W_k 为 a_k 表达式中作为分母的和式.

应用

关于离散数据的最小二乘多项式有二个最主要的应用.

1. **数据平滑化**. 通过接受多项式

$$p(x) = a_0 + a_1 x + \cdots + a_m x^m$$

来替代已知的 $y(x)$, 我们得到一段光滑的线, 如抛物线, 或是其他曲线以替代原先的, 可能不规则的数据函数. $p(x)$ 应该是多少次取决于不同的情况. 通常使用一个 5 点的最小二乘抛物线. 相应于点 (x_i, y_i) 取 $i = k-2, k-1, \cdots, k+2$, 它导出平滑化公式

$$y(x_k) \approx p(x_k) = y_k - \frac{3}{35}\delta^4 y_k,$$

该公式将 5 个值 $y_{k-2}, \cdots, y_{k+2}$ 渗合在一起来为未知的精确值 $y(x_k)$ 提供一个新的估计. 在有限数据表的两端点附近要求稍加修正.

一组近似值 A_i 对相应的真值 T_i 的**标准差**定义为

$$\text{标准差} = \left[\sum_{i=0}^{N} \frac{(T_i - A_i)^2}{N}\right]^{1/2}.$$

在 T_i 为已知的各种试验情况下, 我们将用这个误差度量来估计最小二乘平滑化的有效性.

2. **近似微分**. 正如我们早些时见过的那样, 对不规则数据以一个配置多项式进行拟合时会导致导数的十分差的估计. 甚至在数据中的小的误差也会被放大到恼人的程度. 但是一个最小二乘多项式并非进行配置. 它在数据值之间穿过并且提供光滑性. 这个较为光滑的函数通常带给我们对导数(即 $p'(x)$ 的值)较好的估计. 刚才提到的 5 点抛物线导出公式

$$y'(x_k) \approx p'(x_k) = \frac{1}{10h}(-2y_{k-2} - y_{k-1} + y_{k+1} + 2y_{k+2}).$$

在有限数据表的端点附近, 它也要求修正. 这种公式所产生的结果通常比由微分配置多项式所得来的那些结果优秀得多. 然而, 在力求估计 $y''(x_k)$ 时, 对 $p'(x_k)$ 值重复应用它则又会导致有疑问的精度.

连续数据

对于连续数据 $y(x)$ 我们可以极小化积分

$$I = \int_{-1}^{1} [y(x) - a_0 P_0(x) - \cdots - a_m P_m(x)]^2 dx,$$

$P_j(x)$ 为 Legendre 多项式. [我们必须假设 $y(x)$ 为可积的.] 这意味着我们从一开始就选择了正交多项式来表示我们的最小二乘多项式, 其形式为

$$p(x) = a_0 P_0(x) + \cdots + a_m P_m(x),$$

它的系数证明为

$$a_k = \frac{2k+1}{2}\int_{-1}^{1} y(x) P_k(x) dx.$$

为使用 Legendre 多项式方便起见，首先将数据 $y(x)$ 在其上给出的区间规格化为 $(-1,1)$. 有时，使用区间 $(0,1)$ 更为方便. 此时 Legendre 多项式也必须服从自变量的变化. 新的多项式称为**平移的 Legendre 多项式**.

当 $y(x)$ 是复杂结构时，往往某种类型的离散化是必要的. 或者给出系数的积分必须用近似方法加以计算，或者连续变量的集合必须在一开始时就被离散并且要被极小化的是一个和而不是积分. 很清楚，有几种不同的手段好用而计算机必须决定对一个特殊的问题采用那一种手段.

对已知连续数据函数 $y(x)$ 的平滑化和近似微分再次是我们最小二乘多项式 $p(x)$ 的优先应用. 我们简单地接受 $p(x)$ 及 $p'(x)$ 为较不规则的 $y(x)$ 及 $y'(x)$ 的替代量.

最小二乘原则的一个推广包含对积分

$$I=\int_a^b w(x)[y(x)-a_0Q_0(x)-\cdots-a_mQ_m(x)]^2\mathrm{d}x$$

进行极小化，其中 $w(x)$ 是一个非负的权函数. $Q_k(x)$ 为广义的正交函数

$$\int_a^b w(x)Q_j(x)Q_k(x)\mathrm{d}x=0,$$

当 $j\neq k$ 时. 其细节平行于那些已经提到过的 $w(x)=1$ 的情况，系数 a_k 由

$$a_k=\frac{\int_a^b w(x)y(x)Q_k(x)\mathrm{d}x}{\int_a^b w(x)Q_k^2(x)\mathrm{d}x}$$

给出. I 的极小值可以表示为

$$I_{\min}=\int_a^b w(x)y^2(x)\mathrm{d}x-\sum_{k=0}^m W_ka_k^2,$$

其中 W_k 是 a_k 表达式中分母积分. 这就导出 **Bessel 不等式**

$$\sum_{k=0}^m W_ka_k^2\leqslant\int_a^b w(x)y^2(x)\mathrm{d}x$$

以及 m 趋于无穷时级数 $\sum_{k=0}^{\infty}W_ka_k^2$ 为收敛的事实. 假如所涉及的正交族有一种被称作**完备性**的性质和假设 $y(x)$ 为充分光滑的，则级数真正收敛于在 $I_{\min}$ 中出现的积分. 这意味着逼近误差当 $p(x)$ 的次数增加时趋于零.

Chebyshev 多项式

用 Chebyshev 多项式逼近是广义最小二乘法的重要特殊情况 $w(x)=1/\sqrt{1-x^2}$. 积分区间被规格化为 $(-1,1)$，在这种情况下正交多项式 $Q_k(x)$ 为 Chebyshev 多项式

$$T_k(x)=\cos(k\arccos x),$$

其头几项可证明为 $T_0(x)=1$, $T_1(x)=x$, $T_2(x)=2x^2-1$, $T_3(x)=4x^3-3x$.

Chebyshev 多项式的性质包括

$$T_{n+1}(x)=2xT_n(x)-T_{n-1}(x),$$

$$\int_{-1}^1\frac{T_m(x)T_n(x)}{\sqrt{1-x^2}}\mathrm{d}x=\begin{cases}0, & 若\ m\neq n,\\ \pi/2, & 若\ m=n\neq 0,\\ \pi, & 若\ m=n=0.\end{cases}$$

$$T_n(x)=0\quad 当\ x=\cos[(2i+1)\pi/2n],\ i=0,1,\cdots,n-1,$$

$$T_n(x)=(-1)^i\quad 当\ x=\cos(i\pi/n),\ i=0,1,\cdots,n.$$

一个特别吸引人的性质是等误差性质，它与 Chebyshev 多项式在极值 ±1 之间振动有关，在区间 $(-1,1)$ 内的 $n+1$ 个点处达到这些极值. 作为这一性质的一个推论是，误差 $y(x)-p(x)$ 通常被发现在近似地为 $\pm E$ 的极大值与极小值之间振动. 这样一种几乎等误差

止是所期望的,因为它隐含了我们的逼近经过整个区间时有几乎均匀的精度.关于精确的等误差性质见下一章.

经过简单的处理,x 的幂可以用 Chebyshev 多项式来展示,例如,

$$1 = T_0,\quad x = T_1,\quad x^2 = \frac{1}{2}(T_0 + T_2),\quad x^3 = \frac{1}{4}(3T_1 + T_3).$$

这提示了一个称作多项式减缩的过程,在一个多项式中的每个 x 的幂凭借它都可以用 Chebyshev 多项式的相应组合来替代.经常发现许多高次 Chebyshev 多项式因而会降低,然后保留下来的项组成一个对原多项式的最小二乘逼近,以足够的精度来迎合多种意图.所获结果将会有几乎等误差性质.这个减缩过程可以近似地替代用 Chebyshev 多项式逼近时对系数积分的直接估算.令人不愉快的权因子 $w(x)$ 使得这些积分对大多数的 $y(x)$ 而言令人望而却步.

最小二乘原则的另一变种是将和

$$\sum_{i=0}^{N-1}[y(x_i) - a_0T_0(x_i) - \cdots - a_mT_m(x_i)]^2$$

进行极小化,自变量 $x_i = \cos[(2i+1)\pi/2N]$.这些自变量可以被识别为 $T_N(x)$ 的零点.系数容易用 Chebyshev 多项式的第二个正交性质来决定.

$$\sum_{i=0}^{N-1} T_m(x_i)T_n(x_i) = \begin{cases} 0, & 若\ m \neq 0, \\ N/2, & 若\ m = n \neq 0, \\ N, & 若\ m = n = 0. \end{cases}$$

并证明为 $a_0 = \frac{1}{N}\sum_{i=0}^{N-1} y(x_i),\quad a_k = \frac{2}{N}\sum_{i=0}^{N-1} y(x_i)T_k(x_i).$

于是逼近多项式自然为

$$p(x) = a_0T_0(x) + \cdots + a_mT_m(x).$$

这个多项式也有一个几乎等误差性质.

L_2(模)范数

本章的根本主题是极小化范数

$$\| y - p \|_2,$$

其中 y 表示给定的数据而 p 表示逼近多项式.

题　解

离散数据,最小二乘直线

21.1 找出直线 $p(x) = Mx + B$,对它来说 $\sum_{i=0}^{N}(y_i - Mx_i - B)^2$ 为极小,数据 (x_i, y_i) 为给定的.

解 ☞ 记该和为 S,我们按照一个寻找极小的标准方案令导数为零.

$$\frac{\partial S}{\partial B} = -2\sum_{i=0}^{N} 1 \cdot (y_i - Mx_i - B) = 0,$$

$$\frac{\partial S}{\partial M} = -2\sum_{i=0}^{N} x_i \cdot (y_i - Mx_i - B) = 0.$$

经改写我们得到

$$(N+1)B + \left(\sum x_i\right)M = \sum y_i,$$

$$\left(\sum x_i\right)B + \left(\sum x_i^2\right)M = \sum x_iy_i.$$

它们就是"法方程".引进符号

$$s_0 = N+1,\quad s_1 = \sum x_i,\quad s_2 = \sum x_i^2,\quad t_0 = \sum y_i,\quad t_1 = \sum x_iy_i.$$

这些方程可以解成

$$M=\frac{s_0t_1-s_1t_0}{s_0s_2-s_1^2},\qquad B=\frac{s_2t_0-s_1t_1}{s_0s_2-s_1^2}.$$

的形式，为了证明 $s_0s_2-s_1^2$ 不为零，我们首先注意平方并相加诸如 $(x_0-x_1)^2$ 那样的项导出

$$0<\sum_{i<j}(x_i-x_j)^2=N\cdot\sum x_i^2-2\sum_{i<j}x_ix_j.$$

还有

$$\left(\sum x_i\right)^2=\sum x_i^2+2\sum_{i<j}x_ix_j.$$

所以 $s_0s_2-s_1^2$ 变为

$$(N+1)\sum x_i^2-\left(\sum x_i\right)^2=N\cdot\sum x_i^2-2\sum_{i<j}x_ix_j>0.$$

此处我们假设了 x_i 不全相同，这肯定是合理的．最后一个不等式还帮助我们去证明所选的 M 及 B 真的产生了一个极小．计算二阶导数，我们得到

$$\frac{\partial^2S}{\partial B^2}=2s_0\quad\frac{\partial^2S}{\partial M^2}=2s_2\quad\frac{\partial^2S}{\partial B\partial M}=2s_1.$$

由于头二个为正的且由于

$$(2s_1)^2-2(N+1)2s_2=4(s_1^2-s_0s_2)<0,$$

所以关于二变量 B 及 M 的函数之极小点的二阶导数检验被满足．一阶导数只有一次同时为零的事实说明我们的极小是一个绝对极小．

21.2 赋予各种障碍的高尔夫球员们在一个困难的标准杆数为 3 的球洞处记录的平均成绩如下：

障碍数	6	8	10	12	14	16	18	20	22	24
平均	3.8	3.7	4.0	3.9	4.3	4.2	4.2	4.4	4.5	4.5

用题 21.1 的方式对这个数据求出最小二乘线性函数．

解 ☞ 令 h 表示障碍数和 $x=(h-6)/2$，则 x_i 为整数 $0,\cdots,9$．令 y 表示平均成绩．于是 $s_0=10$，$s_1=45$，$s_2=285$，$t_0=41.5$，$t_1=194.1$ 以及

$$M=\frac{(10)(194.1)-(45)(41.5)}{(10)(285)-(45)^2}\approx0.089,$$

$$B=\frac{(285)(41.5)-(45)(194.1)}{(10)(285)-(45)^2}\approx3.76,$$

它使 $y\approx p(x)$，其中 $p(x)=0.09x+3.76\approx0.045h+3.49$．

21.3 用上题中的最小二乘直线平滑被记录的数据．

解 ☞ 可以设想被记录的数据都含有一定程度的不精确性，有理由予以校正，为此进行平滑数据的努力．在这种情况下数据似乎大致落在一根直线周围，但是有大的起伏，或许是由于高尔夫游戏本身就有起伏．（参看下面的图 21.1.）

就描述在障碍数与平均分之间真实关系而言可以假设为，最小二乘直线比起原始数据来是更好表达式．它提供下面的平滑过的值：

障碍数	6	8	10	12	14	16	18	20	22	24
光滑过的 y	3.76	3.85	3.94	4.03	4.12	4.21	4.30	4.39	4.48	4.57

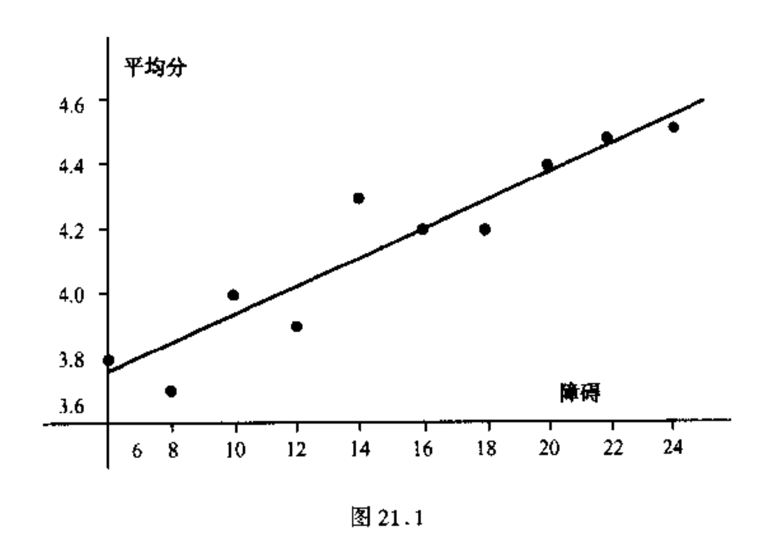


图 21.1

21.4　估计每单位障碍的平均记分增加率.

解　从题 21.2 的最小二乘直线我们得到估计:每单位障碍 0.45 击.

21.5　从下面的数据求得一个具有 $P(x)=Ae^{Mx}$ 型的公式:

x_i	1	2	3	4
P_i	7	11	17	27

解　令 $y=\log P$, $B=\log A$. 然后取对数, $\log P=\log A+Mx$, 它等价于 $y(x)=Mx+B$.

现在我们决定使它作为对 (x_i, y_i) 数据点的最小二乘直线.

x_i	1	2	3	4
y_i	1.95	2.40	2.83	3.30

因为 $s_0=4$, $s_1=10$, $s_2=30$, $t_0=10.48$, $t_1=28.44$, 故由题 21.1 的公式得 $M\approx 0.45$ 及 $B\approx 1.5$. 最终的公式是 $P=4.48e^{0.45x}$.

必须指出在这个过程中我们不是将 $\sum[P(x_i)-P_i]^2$ 极小化, 而是代之以选择对 $\sum[y(x_i)-y_i]^2$ 进行极小化更为简单的尝试. 这是在这类问题中的一个十分普遍的策略.

离散数据,最小二乘多项式

21.6　推广题 21.1, 找出多项式 $p(x)=a_0+a_1x\cdots+a_mx^m$. 对它来说 $S=\sum_{i=0}^{N}(y_i-a_0-a_1x_1-\cdots-a_mx_i^m)^2$ 是一个极小, 数据 (x_i, y_i) 为给定的, 且 $m<N$.

解　如像直线的简单情况那样进行, 令它对 $a_0, a_1, \cdots, a_m$ 的导数为零就产生 $m+1$ 个方程

$$\frac{\partial s}{\partial a_k}=-2\sum_{i=0}^{N}x_i^k(y_i-a_0-a_1x_i-\cdots\quad a_mx_i^m)=0,$$

其中 $k=0,\cdots,m$. 引入符号 $S_k=\sum_{i=0}^{N}x_i^k$, $t_k=\sum_{i=0}^{N}y_ix_i^k$, 这些方程可以改写成

$$s_0a_0+s_1a_1+\cdots+s_ma_m=t_0,$$
$$s_1a_0+s_2a_1+\cdots+s_{m-1}a_m=t_1,$$
$$\cdots$$
$$s_ma_0+s_{m+1}a_1+\cdots+s_{2m}a_m=t_m.$$

称作法方程. 对系数 a_i 解这个方程, 我们得到最小二乘多项式. 我们将证明它只有一解并且它的确使 S 为极小. 对于较小的 m, 这些方程可以没有困难地解出. 对于较大的 m 这个方程组是病态的(坏条件的), 因为另一种处理过程将被提出.

21.7 证明最小二乘思想(正如刚刚在题 21.6 中以及更早些时在题 21.1 中都已被提出)可以推广到任意的向量空间. 与正交投影的关系是什么?

证 ☞ 这个更一般的处理将对稍后在本章中出现的其他最小二乘变种起示范作用, 并且把注意力集中在所有这些变种所具有的共性上. 首先回顾在欧氏平面几何中, 给出一个点 y 和一根直线 S, 在 S 上惟一离 y 最近的点 p 使得 $\overline{py}$ 正交于 S, p 是 y 在 S 上的正交投影点. 类似地在欧氏立体几何中, 给出一个点 y 和一个平面 S, 在平面 S 上惟一离 y 最近的点 p 使得 $\overline{py}$ 正交 S 中的所有向量, p 仍是 y 的正交投影. 这一思想现在扩展到一个更为一般的向量空间.

在向量空间 E 中给出一个向量 y, 要在给定的子空间 S 中寻找一个向量 p 使得

$$\|y-p\| < \|y-q\|,$$

其中 q 是 S 中任何一个其他向量并且向量 v 的模为

$$\|v\| = \sqrt{(v,v)},$$

圆括号表示与向量空间相关联的标量积. 我们从证明有一个惟一的向量 p 开始, 对它来说 $y-p$ 正交于 S 中的每个向量, 这个 p 称作 y 的正交投影.

令 $e_0, \cdots, e_m$ 为对于 S 的正交基并考虑向量

$$p = (y,e_0)e_0 + (y,e_1)e_1 + \cdots + (y,e_m)e_m,$$

直接推算表明对 $k=0,\cdots,m$, $(p,e_k)=(y,e_k)$ 且因此 $(p-y,e_k)=0$. 于是对 S 中的任何 q, 简单地通过以正交基表示 q, 就有 $(p-y,q)=0$. 假如另一个向量 p' 也有这种性质 $(p'-y,q)=0$, 则会得出对 S 中的任何 q 会有 $(p-p',q)=0$. 由于 $p-p'$ 本身在 S 中, 这就迫使 $(p-p',p-p')=0$, 依据对任何标量积都要有的性质, 就隐含了 $p=p'$. 因此正交投影 p 是惟一的.

但是现在, 假若 q 是 S 中不同于 p 的另一个向量,

$$\begin{aligned}\|y-q\|^2 &= \|(y-p)+(p-q)\|^2\\ &= \|y-p\|^2 + \|p-q\|^2 + 2(y-p,p-q).\end{aligned}$$

由于 $p-q$ 在 S 中故最后一项为零, 我们推导出 $\|y-p\| < \|y-q\|$ 正如所求.

21.8 若 $u_0, u_1, \cdots, u_m$ 是 S 的任意一个基, 决定上题中的向量 p 用 u_k 来表示.

解 ☞ 我们必定有 $(y-p,u_k)=0$ 或 $(p,u_k)=(y,u_k)$ 对 $k=0,\cdots,m$. 由于 p 有惟一的表达式

$$p = a_0u_0 + a_1u_1 + \cdots + a_mu_m,$$

代入后直接导出

$$(u_0,u_k)a_0 + (u_1,u_k)a_1 + \cdots + (u_m,u_k)a_m = (y,u_k),$$

对 $k=0,\cdots,m$. 这些是对所给问题的法方程, 要对系数 $a_0,\cdots,a_m$ 来解它. 有惟一的解为前题所保证. 注意在特殊情况下, 当 $u_0,\cdots,u_m$ 为正交规范基时, 这些法方程如在题 21.7 给出的证明中那样, 简化为 $a_i=(y,u_i)$.

还要注意下面重要的推论. 若 y 本身是以 E 中的正交基所表示, 其中包含 $u_0,\cdots,u_m$, 譬如说

$$y = a_0u_0 + a_1u_1 + \cdots + a_mu_m + a_{m+1}u_{m+1} + \cdots,$$

则作为最小二乘近似的正交投影 p, 只要简单地把表达式在 a_mu_m 处予以截断就可以了:

$$p = a_0u_0 + a_1u_1 + \cdots + a_mu_m.$$

21.9 在题 21.6 中处理过的特定情况与在题 21.7 及 21.8 中给出的推广是如何联系的?

解 ☞ 必须作下面的标识:

E: 在变量 $x_0,\cdots,x_N$ 集合上的离散实值函数的空间.

S: E 的包含小于等于 m 次多项式的子集.

y: 具有值 $y_0,\cdots,y_N$ 的数据函数.

(v_1,v_2): 标量积 $\sum_{i=0}^{N} v_1(x_i)v_2(x_i)$.

$\|v\|^2$: 范数 $\sum_{i=0}^{N} [v(x_i)]^2$.

u_k:具有值 x_i^k 的函数.

p:具有值 $p_i = a_0 + a_1 x_i + \cdots + a_m x_i^m$ 的多项式.

$\|y - p\|^2$:和 $S = \sum_{i=0}^{N}(y_i - p_i)^2$.

(y, u_k): $t_k = \sum_{i=0}^{N} y_i x_i^k$.

(u_j, u_k): $S_{j+k} = \sum_{i=0}^{N} x_i^{j+k}$.

以这些标识符我们还认识到题 21.6 的多项式 p 是惟一的并且确实提供了最小和. 题 21.7 及 21.8 的一般结果确立了它.

21.10 对题 21.2 的数据决定最小二乘二次函数.

解 ☞ 和数 s_0, s_1, s_2, t_0 及 t_1 已经计算过. 我们还需要 $s_3 = 2025$, $s_4 = 15,333$, 以及 $t_2 = 1292.9$, 这些使得法方程可写成

$$10a_0 + 45a_1 + 285a_2 = 41.5, \qquad 45a_0 + 285a_1 + 2025a_2 = 194.1,$$
$$285a_0 + 2025a_1 + 15,333a_2 = 1248.$$

在一番工作之后它们产生 $a_0 = 3.73$, $a_1 = 0.11$ 和 $a_2 = -0.0023$, 因而我们的二次函数为

$$p(x) = 3.73 + 0.11x - 0.0023x^2.$$

21.11 应用上题的二次函数来平滑记录数据.

解 ☞ 假设这些数据是我们二次函数的值, 我们得到这些值:

障碍	6	8	10	12	14	16	18	20	22	24
经过光滑的 y	3.73	3.84	3.94	4.04	4.13	4.22	4.31	4.39	4.46	4.53

这些与直线假设的预测相差无几, 而对应二次函数的抛物线不会与图 21-1 的直线有显著的差异. a_2 是如此地小的事实已经说明了在高尔夫问题中二次的假设可能是不必要的.

平滑化与微分

21.12 导出关于 5 个点 (x_i, y_i) 的最小二乘抛物线的公式, 其中 $i = k-2, k-1, k, k+1, k+2$.

解 ☞ 令抛物线为 $p(t) = a_0 + a_1 t + a_2 t^2$, 其中 $t = (x - x_k)/h$, 假设自变量 x_i 为等距分布, 其间隔为 h. 被涉及的 5 个点现在的自变量是 $t = -2, -1, 0, 1, 2$. 对这种对称的排列, 法方程简化为

$$5a_0 \qquad + 10a_2 = \sum y_i,$$
$$10a_1 \qquad = \sum t_i y_i,$$
$$10a_0 \qquad + 34a_2 = \sum t_i^2 y_i.$$

因而容易解出. 首先我们得到

$$\begin{aligned} 70a_0 &= 34\sum y_i - 10\sum t_i^2 y_i \\ &= -6y_{k-2} + 24y_{k-1} + 34y_k + 24y_{k-1} - 6y_{k+2} \\ &= 70y_k - 6(y_{k-2} - 4y_{k-1} + 6y_k - 4y_{k+1} + y_{k+2}). \end{aligned}$$

由此得

$$a_0 = y_k - \frac{3}{35}\delta^4 y_k.$$

回代后我们还可得

$$a_2 = \frac{1}{14}(2y_{k-2} - y_{k-1} - 2y_k - y_{k+1} + 2y_{k+2}).$$

从中间的方程直接得

$$a_1 = \frac{1}{10}(-2y_{k-2} - y_{k-1} + y_{k-1} + 2y_{k+2}).$$

21.13 以 $y(x_k)$ 表示精确值，y_k 是它的一个近似值，导出平滑化公式 $y(x_k)\approx y_k+\frac{3}{35}\delta^4 y_k$.

解 ☞ 关于 5 个点 (x_{k-2},y_{k-2}) 到 (x_{k+2},y_{k+2}) 的最小二乘抛物线为

$$p(x)=a_0+a_1t+a_2t^2.$$

在当中一个自变量 $t=0$ 处由题 21.12 它变成 $p(x_k)=a_0=y_k-\frac{3}{35}\delta^4 y_k$，使用这个公式相当于接受在抛物线上的 p 值作为比数据值 y_k 更好的值.

21.14 从 1 到 10 的整数的平方根被舍入到二位小数，并在每个数上加上一个值为 -0.05，$-0.04\cdots,0.05$ 之一的随机误差（从标有这些数的 11 张卡片中随机抽出而确定），其结果形成表 21.1 中的顶行. 利用上题中的公式平滑这些数值.

解 ☞

表 21.1

x_k	1		2		3		4		5		6		7		8		9		10
y_k	1.04		1.37		1.70		2.00		2.26		2.42		2.70		2.78		3.00		3.14
δy		33		33		30		26		16		28		8		22		14	
$\delta^2 y$			0		−3		−4		−10		12		−20		14		−8		
$\delta^3 y$				−3		−1		−6		22		−32		34		−22			
$\delta^4 y$					2		−5		28		−54		66		−56				
$\frac{3}{35}\delta^4 y$					0		0		2		−5		6		−5				
$p(x_k)$					1.70		2.00		2.24		2.47		2.64		2.83				

直到 4 阶的差分以及 $\frac{3}{35}\delta^4 y$ 也出现在表 21.1 中. 最后，底行包含了平滑后的值.

21.15 题 21.13 的平滑化公式要求在 x_k 的二侧各二个数据值产生平滑过的值 $p(x_k)$，因此它不能用在数据表的头二个及最后二个成员上. 导出用于平滑端点值的公式

$$y(x_0)\approx y_0+\frac{1}{5}\Delta^3 y_0+\frac{3}{35}\Delta^4 y_0,\qquad y(x_{N-1})\approx y_{N-1}+\frac{2}{5}\nabla^3 y_N-\frac{1}{7}\nabla^4 y_N,$$

$$y(x_1)\approx y_1-\frac{2}{5}\Delta^3 y_0-\frac{1}{7}\Delta^4 y_0,\qquad y(x_N)\approx y_N-\frac{1}{5}\nabla^3 y_N+\frac{3}{35}\nabla^4 y_N.$$

解 ☞ 假如我们令 $t=(x-x_2)/h$，则题 21.12 的二次函数就是对头 5 点的最小二乘二次函数. 我们将用这个函数在 x_0 及 x_1 处的值作为 y 的平滑值. 首先

$$p(x_0)=a_0-2a_1+4a_2,$$

并插入关于 a_i 的表达式（以 2 来代替 k），得

$$\begin{aligned}p(x_0)&=\frac{1}{70}(62y_0+18y_1-6y_2-10y_3+6y_4)\\&=y_0+\frac{1}{70}[(-14y_0+42y_1-42y_2+14y_3)\\&\quad+(6y_0-24y_1+36y_2-24y_3+6y_4)],\end{aligned}$$

这就回到上面关于 $y(x_0)$ 的公式. 对 $p(x_1)$ 我们有

$$p(x_1)=a_0-a_1+a_2,$$

并再一次将关于 a_i 的表达式插入而导出所要求的公式. 在我们数据表的另一端，应用变量变换 $t=(x-x_{N-2})/h$，其细节类似.

21.16 应用上题的公式来完成表 21.1 中 y 值的平滑化.

解 ☞ 我们发现它们改变到二位.

$$y(x_0)\approx 1.04+\frac{1}{5}(-0.03)+\frac{3}{35}(0.02)\approx 1.03,$$

$$y(x_{N-1})\approx 3.00+\frac{2}{5}(-0.22)-\frac{1}{7}(-0.56)\approx 2.99,$$

$$y(x_1)\approx 1.37-\frac{2}{5}(-0.03)-\frac{1}{7}(0.02)\approx 1.38,$$

$$y(x_N)\approx 3.14-\frac{1}{5}(-0.22)+\frac{3}{35}(-0.56)\approx 3.14.$$

21.17 对原始数据与对平滑过的值计算标准差.

解 ☞ 对应于精确值 T_i 的一组近似值 A_i 的标准差定义为

$$\text{标准差} = \left[\sum_{i=0}^{N}\frac{(T_i - A_i)^2}{N}\right]^{1/2}.$$

在这个例子中我们有下面的值:

T_i	1.00	1.41	1.73	2.00	2.24	2.45	2.65	2.83	3.00	3.16
y_i	1.04	1.37	1.70	2.00	2.26	2.42	2.70	2.78	3.00	3.14
$p(x_i)$	1.03	1.38	1.70	2.00	2.24	2.47	2.64	2.83	2.99	3.14

精确方根值以二位小数给出.依据上面的公式,

$$y_i\text{ 的标准差}\approx\left(\frac{0.0108}{10}\right)^{1/2}\approx 0.033,$$

$$p(x_i)\text{ 的标准差}\approx\left(\frac{0.0037}{10}\right)^{1/2}\approx 0.019.$$

所以误差减少几近一半.中心部分改进更大一些.如果我们在每一端略去二个值则我们得到它们的标准差分别为 0.035 及 0.015,减少了比一半还多.题 21.13 的公式显得比题 21.15 的那些更为有效.

21.18 使用 5 点抛物线来得到关于近似微分的公式

$$y'(x_k)\approx\frac{1}{10h}(-2y_{k-2}-y_{k-1}+y_{k+1}+2y_{k+2})$$

解 ☞ 用题 21.13 的符号我们将使用 $y'(x_k)$,它是我们 5 点抛物线的导数,作为对在 x_k 处精确导数的一个近似.这又相当于假设我们的数据值 y_i 为精确的,但又是未知的函数之近似值,然而 5 点抛物线将是一个更好的近似,特别是在中心点附近.在抛物线

$$p = a_0 + a_1 t + a_2 t^2$$

上并且按照计划,我们计算 $p'(t)$在 $t=0$ 处的值,它就是 a_1.要把它转化为对 x 的导数仅涉及除以 h 的除法,于是重现在题 21.12 中所获得的 a_1 并取 $p'(x)$为$y'(x)$的一个近似值,我们便得到所要求的公式.

21.19 应用上面的公式由表 21.1 中所给的 y_k 值来估计$y'(x)$.

解 ☞ 在 $x_2=3$ 处我们得到

$$y'(3)\approx\frac{1}{10}(-2.08-1.37+2.00+4.52)=0.307.$$

而在 $x_3=4$ 处

$$y'(4)\approx\frac{1}{10}(-2.74-1.70+2.26+4.84)=0.266.$$

顶行所列的其他成员是以同样的方式得到的.第二行用早些从 Stirling 5 点配置多项式得到的近似公式

$$y'(x_k)\approx\frac{1}{12h}(y_{k-2}-8y_{k-1}+8y_{k+1}-y_{k+2})$$

进行计算.注意现在这个公式的优越性.早些时发现数据中的误差被近似微分公式放大到可观的程度.通过减少此种数据误差的预平滑能带来更好的结果.

由最小二乘所得 $y'(x)$	0.31	0.27	0.24	0.20	0.18	0.17
由配置所得 $y'(x)$	0.31	0.29	0.20	0.23	0.18	0.14
准确的 $y'(x)$	0.29	0.25	0.22	0.20	0.19	0.18

21.20　题 21.18 公式不用于数据表的两端.在每一端用一个 4 点抛物线公式来得到公式

$$y'(x_0) \approx \frac{1}{20h}(-21y_0 + 13y_0 + 17y_2 - 9y_3),$$

$$y'(x_1) \approx \frac{1}{20h}(-11y_0 + 3y_1 + 7y_2 + y_3),$$

$$y'(x_{N-1}) \approx \frac{1}{20h}(11y_N - 3y_{N-1} - 7y_{N-2} \quad y_{N-3}),$$

$$y'(x_N) \approx \frac{1}{20h}(21y_N - 13y_{N-1} - 17y_{N-2} + 9y_{N-3}).$$

解　用 4 点而不用 5 点,其想法是第 5 点可能会离要求导数的位置 x_0 或 x_N 相当远.依据 h 的大小和数据的光滑程度,或许还有其他的因素,人们可能使用基于 5 点或更多点的公式.现在对 4 点抛物线公式我们令 $t=(x-x_1)_h$,于是头 4 个点有自变量为 $t=-1,0,1,2$. 法方程变成

$$4a_0 + 2a_1 + 6a_2 = y_0 + y_1 + y_2 + y_3,$$

$$2a_0 + 6a_1 + 8a_2 = -y_0 + y_2 + 2y_3,$$

$$6a_0 + 8a_1 + 18a_2 = y_0 + y_2 + 4y_3.$$

而且可以解成

$$20a_0 = 3y_0 + 11y_1 + 9y_2 - 3y_3, \qquad 20a_1 = -11y_0 + 3y_1 + 7y_2 + y_3,$$

$$4a_2 = y_0 - y_1 - y_2 + y_3.$$

用这些以及 $y'(x_0)=(a_1-2a_2)/h$, $y'(x_1)=a_1/h$,
便得所要求的结果,在数据表的另一端其细节几乎是相同的.

21.21　应用上题中的公式于表 21.1 中的数据.

解　我们得到

$$y'(1) \approx \frac{1}{20}[-21(1.04) + 13(1.37) + 17(1.70) - 9(2.00)] \approx 0.35,$$

$$y'(2) \approx \frac{1}{20}[-11(1.04) + 3(1.37) + 7(1.70) + 2.00] \approx 0.33.$$

类似地有 $y'(9)\approx 0.16$ 及 $y'(10)\approx 0.19$. 准确值为 0.50, 0.35, 0.17, 及 0.16. 在两端处所得到的不太好的结果是数值微分难点的进一步的例证. Newton 的原始公式

$$y'(x_0) \approx \Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \frac{1}{4}\Delta^4 y_0 + \cdots.$$

由这个数据产生的值为 0.32,它比我们的 0.35 差.在另外一端相应的后差公式得出的 0.25 比我们的 0.19 更差得多.

21.22　使用表 21.1 中的数据,第二次应用求近似导数的公式来估计 $y''(x)$.

解　我们已经得到一阶导数的估计值,粗略地有二位精度.它们的值如下:

x	1	2	3	4	5	6	7	8	9	10
$y'(x)$	0.35	0.33	0.31	0.27	0.24	0.20	0.18	0.17	0.16	0.19

现在应用相同的公式于 $y'(x)$ 值将产生 $y''(x)$ 的估计.例如,在 $x=5$ 处

$$y''(t) \approx \frac{1}{10}[-2(0.31) + (0.27) + (0.20) + 2(0.18)]$$

$$= -0.033,$$

(误差)再次大到为准确值 -0.022 的一半.从我们的公式所得的全部结果与准确值如下:

y''(计算值)	0.011	0.021	0.028	0.033	0.033	0.026	0.019	0.004	0.012	−0.32
$-y''$(准确值)	0.250	0.088	0.048	0.031	0.022	0.017	0.013	0.011	0.009	0.008

靠近中央部分我们有偶尔的一线希望然而在端点处灾难是显见的.

21.23 对于7点的最小二乘抛物线导出平滑公式

$$y(x_k) \approx y_k - \frac{3}{7}\delta^4 y_k - \frac{2}{21}\delta^6 y_k.$$

(推导过程要作为一个补充题.)应用该公式于表21.1的数据,它是否提供比5点平滑公式更好的值?

解 ☞ 在表21.1中可以加上6阶差分的一行:

$$40 \quad -115 \quad 202 \quad -242$$

于是公式产生

$$y(4) \approx 2.00 - \frac{3}{7}(-0.05) - \frac{2}{21}(0.40) \approx 1.98,$$

$$y(5) \approx 2.26 - \frac{3}{7}(0.28) - \frac{2}{21}(-1.15) \approx 2.25,$$

以及类似地有 $y(6)\approx 2.46$, $y(7)\approx 2.65$. 这些结果比5点公式的略有改进,稍微差一些的 $y(4)$ 除外.

正交多项式,离散情况

21.24 对于大的 N 及 m 法方程组可以是严重病态的.为了明白这一点,可证明对从0到1间等距的 x_i 其系数矩阵假如在每一项中删去一个 N 因子,则近似地为

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{m+1} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{m+2} \\ & & \cdots\cdots\cdots & & \\ \frac{1}{m+1} & \frac{1}{m+2} & \frac{1}{n+3} & \cdots & \frac{1}{2m+1} \end{pmatrix}$$

这矩阵是 $m+1$ 阶 Hilbert 矩阵.

解 ☞ 对于大的 N 在 $y(x)=x^k$ 之下在0与1之间的区域面积近似地为 N 个矩形之和.(见图21.2)由于精确的面积由一个积分给出,所以我们有

$$\frac{1}{N}\sum_{i=0}^{N} x_i^k \approx \int_0^1 x^k dx = \frac{1}{k+1}.$$

因此 $s_k \approx N/(k+1)$,而删去 N 后我们立得 Hilbert 矩阵.稍后将证明该矩阵对于大的 N 特别麻烦.

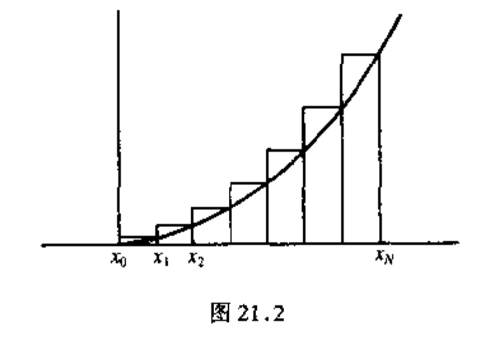


图21.2

21.25 怎样才能避开 Hilbert 矩阵?

解 ☞ 上一题中说明以 $1, x, \cdots, x^m$ 为基底以及变量等距分布所产生的法方程包含一个烦人的近似 Hilbert 矩阵.从计算角度来看更为有效的是找到一个正交基使得相应的法方程变成平凡的.在下题中将构造一个方便的正交基.有趣的是注意到在推导这个基时我们将直接处理 Hilbert 矩阵本身,而不是它的近似,并且出现的方程组将被精确地解出,因而避开了计算病态方程这个陷坑.(同时参看第二十六章.)

21.26 构造一组次数为 $m=0,1,2,\cdots$ 的多项式使得

$$\sum_{t=0}^{N} P_{m,N}(t) P_{n,N}(t) = 0, \qquad 当\ m > n.$$

这类多项式称作在自变量 t 的集合上正交的.

解 令这个多项式是

$$P_{m,N}(t) = 1 + c_1 t + c_2 t^{(2)} + \cdots + c_m t^{(m)},$$

其中 $t^{(i)}$ 是阶乘 $t(t-1)\cdots(t-i+1)$.我们首先使多项式对于 $S=0,1,\cdots,m-1$ 正交于 $(t+s)^{(s)}$,它意味着我们要求

$$\sum_{t=0}^{N} (t+s)^{(s)} P_{m,N}(t) = 0.$$

由于 $(t+s)^{(s)} P_{m,N}(t) = (t+s)^{(s)} + c_1 (t+s)^{(s+1)} + \cdots + c_m (t+s)^{(s+m)}$,

对变量 t 求和并用题 4.10 就得到

$$\sum_{t=0}^{N} (t+s)^{(s)} P_{m,N}(t)$$
$$= \frac{(N+s+1)^{(s+1)}}{s+1} + c_1 \frac{(N+s+1)^{s-2}}{s+2} + \cdots + c_m \frac{(N+s+1)^{(s+m+1)}}{s+m+1},$$

它应该为零.将 $(N+s+1)^{(s+1)}$ 因子移开,这个和就变成了

$$\frac{1}{s+1} + \frac{Nc_1}{s+2} + \frac{N^{(2)}c_2}{s+3} + \cdots + \frac{N^{(m)}c_m}{s+m-1} = 0,$$

并令 $N^{(i)} c_i = a_i$,方程简化为

$$\frac{1}{s+1} + \frac{a_1}{s+2} + \frac{a_2}{s+3} + \cdots + \frac{a_m}{s+m+1} = 0,$$

当 $s=0,1,\cdots,m-1$. Hilbert 矩阵再一次出现在这个方程组中,但是精确地解这个方程组还将带给我们一个有趣的算法.假如最后的和被合并成单个商式,则它具有形式 $Q(s)/(s+m+1)^{(m+1)}$,其中 $Q(s)$ 为一个次数最多为 m 的多项式.由于 $Q(s)$ 必须在 m 个自变量 $s=0,1,\cdots,m-1$ 处为零,故我们必定有 $Q(s)=Cs^{(m)}$,其中 C 与 s 无关.为了决定 C 我们以 $(s+1)$ 乘这个和以及等价的商式,我们有

$$1 + (s+1)\left(\frac{a_1}{s+2} + \cdots + \frac{a_m}{s+m+1}\right) = \frac{Cs^{(m)}}{(s+2)\cdots(s+m+1)},$$

除了分母的零点外它对所有的 s 均成立.令 $s=-1$,我们发现 $C = m!\ /[(-1)(-2)\cdots(-m)] = (-1)^m$.现在我们有

$$\frac{1}{s+1} + \frac{a_1}{s+2} + \cdots + \frac{a_m}{s+m-1} = \frac{(-1)^m s^{(m)}}{(s+m+1)^{(m+1)}}.$$

现在用产生 C 的措施来产生 a_i.乘以 $(s+m+1)^{(m+1)}$ 然后令 $s=-i-1$,对 $i=1,\cdots,m$ 得到

$$(-1)^i i!(m-i)!\, a_i = (-1)^m (-i-1)^{(m)} = (m+i)^{(m)}.$$

接着对 a_i 求解 $a_i = (-1)^i \dfrac{(m+i)^{(m)}}{(m-i)!\ i!} = (-1)^i \binom{m}{i}\binom{m+1}{i}$.

回忆起 $a_i = c_i N^{(i)}$,所要求的多项式可以写成

$$P_{m,N}(t) = \sum_{i=0}^{m} (-1)^i \binom{m}{i}\binom{m+i}{i} \frac{t^{(i)}}{N^{(i)}}.$$

我们已经证明的是每一个 $P_{m,N}(t)$ 正交于函数

$$1 \quad t+1 \quad (t+2)(t+1)\cdots(t+m-1)^{(m-1)}.$$

但是在题 4.18 中,我们已看到幂 $1,t,t^2,\cdots,t^{m-1}$ 可以表示为这些函数的组合,所以 $P_{m,N}(t)$ 也与每个这样的幂正交.最后,由于 $P_{n,N}(t)$ 是这些幂的组合,我们得到 $P_{m,N}(t)$ 与 $P_{n,N}(t)$ 它们自身也是正交的.这些多项式的前五个为

$$P_{0,N} = 1,$$

$$P_{1,N} = 1 - \frac{2t}{N},$$

$$P_{2,N} = 1 - \frac{6t}{N} + \frac{6t(t-1)}{N(N-1)},$$

$$P_{3,N} = 1 - \frac{12t}{N} + \frac{30(t-1)}{N(N-1)} - \frac{20t(t-1)(t-2)}{N(N-1)(N-2)},$$

$$P_{4,N}=1-\frac{20t}{N}-\frac{90t(t-1)}{N(N-1)}-\frac{140t(t-1)(t-2)}{N(N-1)(N-2)}$$
$$+\frac{70t(t-1)(t-2)(t-3)}{N(N-1)(N-2)(N-3)}.$$

21.27 决定系数 a_k 使得

$$p(x)=a_0P_{0,N}(t)+a_1P_{1,N}(t)+\cdots+a_mP_{m,N}(t)$$

[取 $t=(x-x_0)/h$]成为对数据(x_t,y_t) $i=0,1,\cdots,N$ 的 m 次最小二乘多项式.

解 ☞ 我们要将

$$S=\sum_{t=0}^{N}[y_t-a_0P_{0,N}(t)-\cdots-a_mP_{m,N}(t)]^2$$

极小化.令对于 a_k 的导数为零,我们有

$$\frac{\partial S}{\partial a_k}=-2\sum_{t=0}^{N}[y_t-a_0P_{0,N}(t)-\cdots-a_mP_{m,N}(t)]P_{k,N}(t)=0$$

当 $k=0,1,\cdots,m$.但是由于正交性故此处的大多数项为零,只有二项起作用,

$$\sum_{t=0}^{N}[y_t-a_kP_{k,N}(t)]P_{k,N}(t)=0.$$

就此解出 a_k,我们得到

$$a_k=\frac{\sum_{t=0}^{N}y_tP_{k,N}(t)}{\sum_{t=0}^{N}P_{k,N}^2(t)},$$

这是正交函数的一个优点.系数 a_k 不是联在一起的,每一个都出现在单个的法方程中.将 a_k 代入 $p(x)$,我们便得最小二乘多项式.

直接由题 21.7 及 21.8 的一般理论可以得到同样的结果.标识符 $E,S,y,(v_1,v_2)$,及 $\|v\|$ 完全如前,现在我们取 $u_k=P_{k,N}(t)$,故正交投影仍为 $p=a_0u_0+\cdots+a_mu_m$.第 k 个法方程为$(u_k,u_k)a_k=(y,u_k)$并导出我们已经获得的 a_k 的表达式.我们的一般理论,现在也保证我们真正地将 S 极小化了,并且 $p(x)$是惟一的解.一个用二阶导数的论证也可以确立这一点,但是现在并不需要.

21.28 证明 S 的极小值取 $\sum_{t=0}^{N}y_t^2-\sum_{k=0}^{m}W_ka_k^2$ 这样的形式,其中 $W_k=\sum_{t=0}^{N}P_{k,N}^2(t)$.

证 ☞ 将和展开便得

$$S=\sum_{t=0}^{N}y_t^2-2\sum_{t=0}^{N}y_t\sum_{k=0}^{m}a_kP_{k,N}(t)$$
$$+\sum_{t=0}^{N}\sum_{j,k=0}^{m}a_ja_kP_{j,N}(t)P_{k,N}(t),$$

右边的第二项等于 $-2\sum_{k=0}^{N}a_k(W_ka_k)=-2\sum_{k=0}^{m}W_ka_k^2$.最末一项由于正交性除 $j=k$ 外均消失,在这种情况下它就成为 $\sum_{k=0}^{m}W_ka_k^2$.将个别的项全代回,

$$S_{\min}=\sum_{t=0}^{N}y_t^2-\sum_{k=0}^{m}W_ka_k^2.$$

注意当逼近多项式的次数 m 增加时极小值 S 会发生什么情况.由于 S 为非负的,$S_{\min}$中的第一项明显地大于第二项.但是第二项随 m 而增加,误差稳定地减少.当 $m=n$ 时由我们早些时的工作获悉,存在着一个配置多项式,在每个变量 $t=0,1,\cdots,N$ 处等于 y_t,这就将 S 减少到零.

21.29 应用正交函数算法对下面的数据寻找一个 3 次的最小二乘多项式:

x_i	0	1	2	3	4	5	6	7	8	9	10
y_i	1.22	1.41	1.38	1.42	1.48	1.58	1.84	1.79	2.03	2.04	2.17

x_i	11	12	13	14	15	16	17	18	19	20
y_i	2.36	2.30	2.57	2.52	2.85	2.93	3.03	3.07	3.31	3.48

解 ☞ 系数 a_j 直接由上题的公式进行计算、用手算的话，W_k 及 $P_{k,N}(t)$ 都有表而且应该加以利用.虽然我们有"内部信息"三次多项式就可以了，但是稍微走远一些是有教益的.一直到 $m=5$ 我们得到

$$a_0 = 2.2276, a_1 = -1.1099, a_2 = 0.1133, a_3 = 0.0119,$$

$$a_4 = 0.0283, a_5 = -0.0038; \text{并以 } x = t,$$

$$p(x) = 2.2276 - 1.1099P_{1,20} + 0.1133P_{2,20} + 0.0119P_{3,20} + 0.0283P_{4,20} - 0.0038P_{5,20}.$$

由正交函数展开的性质通过对结果的截断我们可得到不同次数的最小二乘逼近.此种从一次到五次多项式的值与原始数据一起均给在下面的表 21.2 中.最后一列列出 $y(x)=(x+50)^3/10^5$ 的值，所用的数据是由这些值再加上一个大小直到 0.10 的随机误差而得到的.我们的目的是通过最小二乘平滑尽我们所能消除尽可能多的误差后，重新回到这个三次式.没有先验知识说明一个三次多项式是我们的所求，在选择我们逼近式时存在着某些困难.幸运的是线性逼近后面的那些结果并没有巨大的不一致.对标准差的计算表明，在当前情况下二次逼近比三次逼近要好.

次数	1	2	3	4	5	原始数据
RMS	0.060	0.014	0.016	0.023	0.023	0.069

表 21.2

x	给定的数据	1	2	3	4	5	校正的结果
0	1.22	1.12	1.231	1.243	1.27	1.27	1.250
1	1.41	1.23	1.308	1.313	1.31	1.31	1.327
2	1.38	1.34	1.389	1.388	1.37	1.38	1.406
3	1.42	1.45	1.473	1.469	1.45	1.45	1.489
4	1.48	1.56	1.561	1.554	1.54	1.54	1.575
5	1.58	1.67	1.652	1.645	1.63	1.63	1.663
6	1.84	1.78	1.747	1.740	1.74	1.73	1.756
7	1.79	1.89	1.845	1.839	1.84	1.84	1.852
8	2.03	2.01	1.947	1.943	1.95	1.95	1.951
9	2.04	2.12	2.053	2.051	2.07	2.07	2.054
10	2.17	2.23	2.162	2.162	2.18	2.18	2.160
11	2.36	2.34	2.275	2.277	2.29	2.29	2.270
12	2.30	2.45	2.391	2.395	2.41	2.41	2.383
13	2.57	2.56	2.511	2.517	2.52	2.52	2.500
14	2.52	2.67	2.635	2.642	2.64	2.64	2.621
15	2.85	2.78	2.762	2.769	2.76	2.76	2.746
16	2.93	2.89	2.892	2.899	2.88	2.88	2.875
17	3.03	3.00	3.027	3.031	3.01	3.01	3.008
18	3.07	3.12	3.164	3.165	3.15	3.15	3.144
19	3.31	3.23	3.306	3.301	3.30	3.30	3.285
20	3.48	3.34	3.451	3.439	3.47	3.47	3.430

连续数据，最小二乘多项式

21.30 决定系数 a_i，使得

$$I = \int_{-1}^{1}[y(x) - a_0P_0(x) - a_1P_1(x) - \cdots - a_mP_m(x)]^2\mathrm{d}x$$

为极小，$P_k(x)$ 为 k 次 Legendre 多项式.

解 ☞ 这儿要极小化的不是一个平方和而是一个积分，而且数据也不再是离散的 y_i 值而是一个连续变量 x 的函数 $y(x)$. 使用 Legendre 多项式是十分方便的. 正像在前节中一样它将把用来决定 a_i 的法方程简化为一个十分简单的方程组. 同时，由于任何多项式均可表示为 Legendre 多项式的组合，所以实际上我们是对连续数据解最小二乘多项式问题. 令通常的导数为零，我们得到

$$\frac{\partial I}{\partial a_k} = -2\int_{-1}^{1}[y(x) - a_0P_0(x) - \cdots - a_mP_m(x)]P_k(x)\mathrm{d}x = 0,$$

当 $k=0,1,\cdots,m$. 凭借这些多项式的正交性，这些方程立刻简化成

$$\int_{-1}^{1}[y(x) - a_kP_k(x)]P_k(x)\mathrm{d}x = 0.$$

每个方程只包含一个 a_k，故

$$a_k = \frac{\int_{-1}^{1}y(x)P_k(x)\mathrm{d}x}{\int_{-1}^{1}P_k^2(x)\mathrm{d}x} = \frac{2k+1}{2}\int_{-1}^{1}y(x)P_k(x)\mathrm{d}x.$$

此处再一次证实用下面这些标识符我们的问题是题 21.7 和题 21.8 的特殊情况：

E：在 $-1\leqslant x\leqslant 1$ 上的实值函数的空间.

S：m 次或低于 m 次的多项式.

y：数据函数 $y(x)$.

(v_1, v_2)：标量积 $\int_{-1}^{1}v_1(x)x_2(x)\mathrm{d}x$.

$\| v \|$：在 $k\int_{-1}^{1}[v-(x)]^2\mathrm{d}x$.

u_k：$P_k(x)$.

p：$a_kP_0(x) + \cdots + a_mP_m(x)$.

a_k：$(y, u_k)/(u_k, u_k)$.

因而这些题保证了我们的解 $p(x)$ 为惟一的，并且的确使积分 I 极小化.

21.31 找出在区间 $(0,1)$ 上以一条直线对 $y(t)=t^2$ 的最小二乘逼近.

解 ☞ 这里我们以一段直线去逼近一段抛物线弧. 首先令 $t=(x+1)/2$ 来使得自变量 x 的区间为 $(-1,1)$. 它使得 $y=(x+1)^2/4$. 由于 $P_0(x)=1$ 及 $P_1(x)=x$，系数 a_0 及 a_1 为

$$a_0 = \frac{1}{2}\int_{-1}^{1}\frac{1}{4}(x+1)^2dx = \frac{1}{3}, \qquad a_1 = \frac{3}{2}\int_{-1}^{1}\frac{1}{4}(x+1)^2xdx = \frac{1}{2}.$$

因而最小二乘直线是 $y = \frac{1}{3}P_0(x) + \frac{1}{2}P_1(x) = \frac{1}{3} + \frac{1}{2}x = t - \frac{1}{6}$.

抛物线弧及直线均表示在图 21.3 中. 在直线上与在抛物线上 y 值的差为 $t^2 - t + \frac{1}{6}$，而它的极值在 $t=0, \frac{1}{2}$ 及 1 处其量为 1/6，−1/12 及 1/6. 因此以直线代替抛物线时所产生的误差在端点处比在区间的中央要稍大些. 这误差可以表示成

$$\frac{1}{4}(x+1)^2 - \frac{1}{3}P_0(x) - \frac{1}{2}P_1(x) = \frac{1}{6}P_2(x).$$

而 $P_2(x)$ 的形状更进一步证实误差的性态.

21.32 找出在区间 $(0,\pi)$ 上以一个抛物线对 $y=\sin t$ 的最小二乘逼近.

解 ☞ 令 $t=\pi(x+1)/2$ 来得到对变量 x 的区间 $(-1,1)$. 于是 $y=\sin[\pi(x+1)/2]$. 系数为

$$a_0 = \frac{1}{2}\int_{-1}^{1}\sin\left[\frac{\pi(x+1)}{2}\right]\mathrm{d}x = \frac{2}{\pi},$$

$$a_1 = \frac{3}{2}\int_{-1}^{1}\sin\left[\frac{\pi(x+1)}{2}\right]x\,dx = 0,$$

$$a_2 = \frac{5}{2}\int_{-1}^{1}\sin\left[\frac{\pi(x+1)}{2}\right]\frac{1}{2}(3x^2-1)\,dx = \frac{10}{\pi}\left(1-\frac{12}{\pi^2}\right).$$

因而抛物线为

$$y = \frac{2}{\pi} + \frac{10}{\pi}\left(1-\frac{12}{\pi^2}\right)\frac{1}{2}(3x^2-1)$$
$$= \frac{2}{\pi} + \frac{10}{\pi}\left(1-\frac{12}{\pi^2}\right)\left[\frac{6}{\pi^2}\left(t-\frac{\pi}{2}\right)^2-\frac{1}{2}\right].$$

抛物线及正弦曲线如图 21.4 所示，稍稍有点失真为了更好地强调逼近是在其上与在其下的特征.

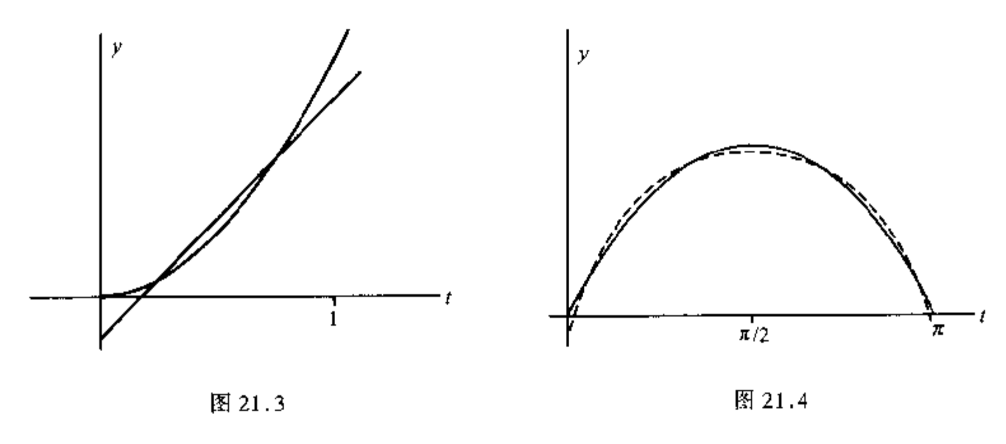


图 21.3　　图 21.4

21.33 什么是"移位的 Legendre 多项式"?

解 这些结果来自于一个变量的变换，它将区间$(-1,1)$转移入$(0,1)$. 令 $t=(1-x)/2$来实现这一变换. 熟悉的以 x 为自变量的 Legendre 多项式于是就变成了以 t 为自变量的：

$$P_0 = 1, \qquad P_2 = \frac{1}{2}(3x^2-1) = 1-6t+6t^2,$$

$$P_1 = x = 1-2t, \quad P_3 = \frac{1}{2}(5x^3-3x) = 1-12t+30t^2-20t^3,$$

等等. 这些多项式是在$(0,1)$上正交的，因而我们可以用它们作为我们连续数据的最小二乘分析的基底以替代标准的 Legendre 多项式. 以这个变量变换，包含在我们公式中关于系数的积分就变成了

$$\int_0^1[P_n(t)]^2dt = \frac{1}{2n+1}, \qquad a_k = (2k+1)\int_0^1 y(t)P_k(t)dt$$

变量变换 $t=(x+1)/2$ 也可以使用，改变每个奇次多项式的符号，然而所采取的措施与题 21.26 中对离散情况的正交多项式所进行的推导非常相似.

21.34 假设一次实验产生的曲线如图 21.5 所示. 已知或是猜想该曲线应为一根直线. 证明最小二乘方直线近似地由 $y=0.21t+0.11$ 给出，在图中以虚线表示.

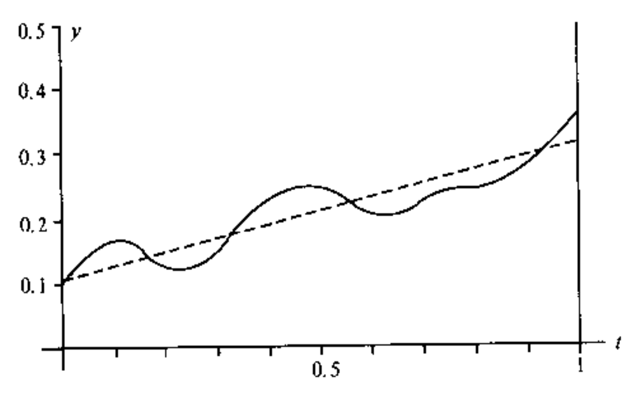


图 21.5

证 ☞ 不去将区间还原到(-1,1),我们直接用变量 t 与平移的 Legendre 多项式进行工作.所需的两个系数为

$$a_0 = \int_0^1 y(t)\mathrm{d}t, \qquad a_1 = 3\int_0^1 y(t)(1-2t)\mathrm{d}t.$$

由于 $y(t)$ 不能用解析的形式,这些积分必须以近似方法加以估值.从图中我们可以估计 y 值如下:

t	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.10
y	0.10	0.17	0.13	0.15	0.23	0.25	0.21	0.22	0.25	0.29	0.36

应用 Simpson 法则现在得到 $a_0 \approx 0.214$ 及 $a_1 \approx -0.105$.结果所获直线为

$$y = 0.214 - 0.105(1-2t) = 0.21t + 0.11$$

它出现在图 21.5 中.解这个问题的另一方法可以包含把关于离散数值的方法应用于从图中读出的 y 值.

连续数据,一种推广的方法

21.35 以一组在区间 (a,b) 上的具有非负数函数 $w(x)$ 的正交多项式来推导最小二乘多项式.

解 ☞ 细节与早些时的那些推导过程十分相似.

我们要通过选择系数 a_k 来极小化

$$I = \int_a^b w(x)[y(x) - a_0Q(x) - \cdots - a_mQ_m(x)]^2\mathrm{d}x,$$

其中函数 $Q_k(x)$ 满足正交条件

$$\int_a^b w(x)Q_j(x)Q_k(x)\mathrm{d}x = 0, \qquad 当\ j \neq k.$$

对涉及导数的完全一样的讨论不存在什么障碍,我们可立刻求助于题 21.7 和 21.8,以标量积

$$(v_1, v_2) = \int_a^b w(x)v_1(x)v_2(x)\mathrm{d}x$$

和其他明显的标识符又可获得

$$a_k = \frac{\int_a^b w(x)y(x)Q_k(x)\mathrm{d}x}{\int_a^b w(x)Q_k^2(x)\mathrm{d}x}$$

具有这些 a_k 的最小二乘多项式为

$$p(x) = a_0Q_0(x) + \cdots + a_mQ_m(x).$$

21.36 a_k 不依赖于 m 的事实有何重要性?

解 ☞ 这意味着逼近多项式的次数不需要在计算一开始时就选定.a_k 可以逐次地决定而且对所需项数的决定可以建立在所计算出的 a_k 的大小上.在非正交性的推理中次数的变化通常要求重新计算所有的系数.

21.37 证明 I 的最小值可以表示为形式

$$\int_a^b w(x)y^2(x)\mathrm{d}x - \sum_{k=0}^m W_k a_k^2 \quad 其中\ W_k = \int_a^b w(x)Q_k^2(x)\mathrm{d}x.$$

证 ☞ 将积分显式地写出就成为

$$I = \int_a^b w(x)y^2(x)\mathrm{d}x - 2\sum_{k=0}^m\int_a^b w(x)y(x)a_kQ_k(x)\mathrm{d}x$$
$$+ \sum_{j,k=0}^m\int_a^b w(x)a_ja_kQ_j(x)Q_k(x)\mathrm{d}x.$$

右侧的第二项等于 $-2\sum_{k=0}^m a_k(W_ka_k) = -2\sum_{k=0}^m W_ka_k^2$.最后一项由于正交性除 $j=k$ 外都等于零,当 j

$=k$ 时它变成 $\sum_{k=0}^{m} W_k a_k^2$. 将各部分一起回代,有 $I_{\min} = \int_a^b w(x) y^2(x) dx - \sum_{k=0}^{m} W_k a_k^2$.

21.38 证明 Bessel 不等式 $\sum_{k=0}^{m} W_k a_k^2 \leqslant \int_a^b w(x) y^2(x) dx$.

证 ☞ 假设 $w(x) \geqslant 0$, 由此得 $I \geqslant 0$ 故 Bessel 不等式是上题的一个直接结果.

21.39 证明级数 $\sum_{k=0}^{\infty} W_k a_k^2$ 为收敛的.

证 ☞ 这是一个正项级数其部分和是,以 Bessel 不等式中的积分为上界的.这保证了收敛.当然,这一直是假设在我们分析中出现的积分是存在的.换言之,我们所处理的函数在区间(a, b)上是可积的.

21.40 当 m 趋于无穷时 $I_{\min}$ 值趋于零是真的吗?

解 ☞ 对通常所用的正交函数族,答案是肯定的.这个过程被称作平均收敛,而这个正交函数集称作是完备的.证明的细节比起将在这里试图作的要更加广泛.

用 Chebyshev 多项式的逼近

21.41 Chebyshev 多项式在 $-1 \leqslant x \leqslant 1$ 中由 $T_n(x) = \cos(n \arccos x)$ 所定义.直接从这个定义找出这类多项式的头几项.

解 ☞ 当 $m=0$ 及 1 时立得 $T_0(x) = 1$, $T_1(x) = x$, 令 $A = \arccos x$, 则

$$T_2(x) = \cos 2A = 2\cos^2 A - 1 = 2x^2 - 1,$$

$$T_3(x) = \cos 3A = 4\cos^3 A - 3\cos A = 4x^3 - 3x, \qquad \text{等等}.$$

21.42 证明递推公式 $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$.

证 ☞ 三角关系式 $\cos(n+1)A + \cos(n-1)A = 2\cos A \cos nA$ 直接转化成 $T_{n+1}(x) + T_{n-1}(x) = 2xT_n(x)$.

21.43 利用递推公式再推出 Chebyshev 多项式的接下去的少数几个成员.

解 ☞ 从 $n=3$ 开始,

$$T_4(x) = 2x(4x^3 - 3x) - (2x^2 - 1) = 8x^4 - 8x^2 + 1,$$

$$T_5(x) = 2x(8x^4 + 1) - (4x^3 - 3x) = 16x^5 - 20x^3 + 5x,$$

$$T_6(x) = 2x(16x^5 - 20x^3 + 5x) - (8x^4 - 8x^2 + 1) = 32x^6 - 48x^4 + 18x^2 - 1,$$

$$T_7(x) = 2x(32x^6 - 48x^4 + 18x^2 - 1) - (16x^5 - 20x^3 + 5x)$$
$$= 64x^7 - 112x^5 + 56x^3 - 7x, \qquad \text{等等}.$$

21.44 证明正交性

$$\int_{-1}^{1} \frac{T_m(x) T_n(x)}{\sqrt{1-x^2}} dx = \begin{cases} 0, & m \neq n, \\ \pi/2, & m = n \neq 0, \\ \pi, & m = n = 0. \end{cases}$$

证 ☞ 令 $x = \cos A$ 如前,上面的积分当 $m \neq n$ 时成为

$$\int_0^{\pi} (\cos mA)(\cos nA) dA = \left[\frac{\sin(m+n)A}{2(m+n)} + \frac{\sin(m-n)A}{2(m-n)} \right]_0^{\pi} = 0.$$

若 $m = n = 0$, 立得结果为 π. 若 $m = n \neq 0$, 积分为

$$\int_0^{\pi} \cos^2 nA dA = \left[\frac{1}{2} \left(\frac{\sin nA \cos nA}{n} + A \right) \right]_0^{\pi} = \frac{\pi}{2}.$$

21.45 以 Chebyshev 多项式表示 x 的幂.

解 ☞ 我们得到

$$1 = T_0, \qquad x = T_1, \qquad x^2 = \frac{1}{2}(T_0 + T_2), \qquad x^3 = \frac{1}{4}(3T_1 + T_3)$$

$$x^4 = \frac{1}{8}(3T_0 + 4T_2 + T_4), \qquad x^5 = \frac{1}{16}(10T_1 + 5T_3 + T_5),$$

$$x^6 = \frac{1}{32}(10T_0 + 15T_2 + 6T_4 + T_6), \qquad x^7 = \frac{1}{64}(35T_1 + 21T_3 + 7T_5 + T_7),$$

等等.很明显,这个过程可以一直持续到任何幂.

21.46　找出将积分

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}}[y(x) - a_0T_0(x) - \cdots - a_mT_m(x)]^2 \mathrm{d}x$$

极小化的最小二乘多项式.

解　凭借前节的结果系数 a_k 为

$$a_k = \frac{\int_{-1}^{1} w(x)y(x)T_k(x)\mathrm{d}x}{\int_{-1}^{1} w(x)T_k^2(x)\mathrm{d}x} = \frac{2}{\pi}\int_{-1}^{1} \frac{y(x)T_k(x)}{\sqrt{1-x^2}}\mathrm{d}x,$$

a_0 除外, $a_0 = \frac{1}{\pi}\int_{-1}^{1} \frac{y(x)}{\sqrt{1-x^2}}\mathrm{d}x$. 最小二乘多项式是 $a_0T_0(x) + \cdots + a_mT_m(x)$.

21.47　证明 $T_n(x)$在区间$(-1,1)$内部有 n 个零点而没有一个在区间之外.什么是“等波纹(equalripple)”性质?

证　由于 $T_n(x) = \cos n\theta$,取 $x = \cos\theta$ 及 $-1 \leqslant x \leqslant 1$,我们可以不失一般性地要求 $0 \leqslant \theta \leqslant \pi$,事实上它使 θ 与 x 之间的关系更加确切.显然当 $\theta = (2i+1)\pi/2n$ 时,或者说当

$$x_i = \cos\frac{(2i+1)\pi}{2n}, \quad i = 0,1,\cdots,n-1$$

时 $T_n(x)$为零.这些是在-1与1之间的 n 个不同的点.因为 $T_n(x)$只有 n 个零点,所以在区间外部就不可能再有零点.由于在区间$(-1,1)$中就等于余弦函数,故在那里 $T_n(x)$在幅度上不会超过1.它在包括端点在内的 $n+1$ 个点处达到这个最大的幅度,

$$T_n(x) = (-1)^i, \qquad \text{在} \quad x_i = \cos\frac{i\pi}{n} \quad \text{处} \quad i = 0,1,\cdots,n.$$

在极值之间的这种等振幅振荡被称作等波纹性.这个性质在图21.6中可以 $T_2(x)$, $T_3(x)$, $T_4(x)$ 及 $T_5(x)$为例证.

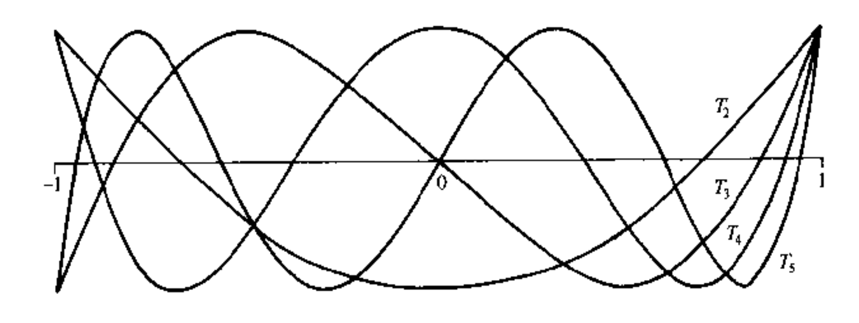


图 21.6

21.48　等波纹性以什么方式使得最小二乘逼近

$$y(x) \approx a_0T_0(x) + \cdots + a_mT_m(x)$$

优于用其他多项式而不是用 $T_k(x)$时所作的类似的逼近?

解　作为前提,我们假设对所涉及的 $y(x)$而言已得到的级数当 m 趋向无限时它收敛于 $y(x)$,还假设它收敛得足够快,以致于

$$y(x) - a_0T_0(x) - \cdots - a_mT_m(x) \approx a_{m+1}T_{m+1}(x).$$

换言之,截断该级数所造成的误差基本上就是略去的第一项.由于 $T_{m+1}(x)$有等波纹性,我们逼近式的误差贯穿整个区间$(-1,1)$时是在 a_{m+1}与$-a_{m+1}$之间起伏.该误差将不会在区间的一部分比起在另一部分来得实质性地大.误差的均匀性可以看作是在积分中采用了不愉快的权因子1/

$\sqrt{1-x^2}$的一种回报.

21.49 使用权函数 $1/\sqrt{1-x^2}$ 在整个区间(0,1)上求关于 $y(t)=t^2$ 的最小二乘直线.

解 变量变换 $t=(x+1)/2$ 将区间转换到变量 x 的区间(-1,1)并且使得 $y=\frac{1}{4}(x^2+2x+1)$. 假如我们首先注意基本结果

$$\int_{-1}^{1}\frac{x^p}{\sqrt{1-x^2}}\mathrm{d}x=\int_0^{\pi}(\cos A)^p\mathrm{d}A=\begin{cases}\pi, & p=0,\\ 0, & p=1,\\ \pi/2, & p=2,\\ 0, & p=3.\end{cases}$$

则系数 a_0 就变成了(参看题 21.46) $a_0=\frac{1}{4}\left(\frac{1}{2}+0+1\right)=\frac{3}{8}$, 并由于 $y(x)T_1(x)$ 等于 $\frac{1}{4}(x^3+2x^2+x)$, 我们便有 $a_1=\frac{1}{4}(0+2+0)=\frac{1}{2}$. 因此, 最小二乘多项式为

$$\frac{3}{8}T_0(x)+\frac{1}{2}T_1(x)=\frac{3}{8}+\frac{1}{2}x.$$

有第二条而且简单得多的途径得到这个结果. 用题 21.45 中的结果,

$$y(x)=\frac{1}{4}\left\{\frac{1}{2}T_0+\frac{1}{2}T_2+2T_1+T_0\right\}$$
$$=\frac{3}{8}T_0+\frac{1}{2}T_1+\frac{1}{8}T_2.$$

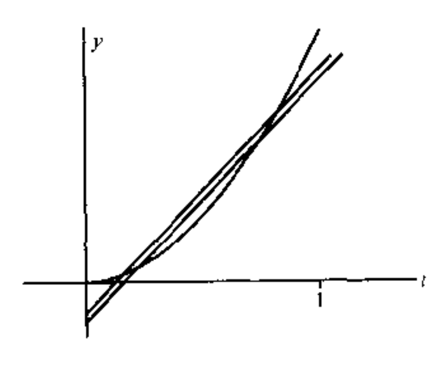


图 21.7

将它从线性项后截断, 我们立得刚获得的结果. 此外, 我们发现在这种二次函数 $y(x)$ 的情况下, 误差正好是等波纹函数 $T_2(x)/8$. 这当然是 Chebyshev 多项式中断于这一项的结果. 对大多数函数而言这个误差将只近似地是第一个被省略的项, 因此只是近似地为一个等波纹误差. 把这儿的极端误差 $\left(\frac{1}{8},-\frac{1}{8},\frac{1}{8}\right)$ 与那些在题 21.31 中为 $\left(\frac{1}{6},-\frac{1}{12},\frac{1}{6}\right)$ 的相比较, 我们发现加以等波纹性眼前的这个逼近式牺牲了在中心部分的某些精度而改进了在极值处的精度. 二条线都画在图 21.7 上.

21.50 求以 Chebyshev 多项式对 $y(x)=\sin x$ 进行三次逼近.

解 为了得到最小二乘多项式的系数所必须计算的带有权函数 $w(x)=1/\sqrt{1-x^2}$ 的积分在这种情况下太复杂了. 代之我们将以例子说明多项式减缩的过程. 从

$$\sin x\approx x-\frac{1}{6}x^3+\frac{1}{120}x^5$$

开始, 使用题 21.45, 我们将 x 的幂代之以它们的用 Chebyshev 多项式表示的等价物.

$$\sin x\approx T_1-\frac{1}{24}(3T_1+T_3)+\frac{1}{1920}(10T_1+5T_3+T_5)$$
$$=\frac{169}{192}T_1-\frac{5}{128}T_3+\frac{1}{1920}T_5.$$

这里的系数不完全是题 21.46 中的 a_k, 因为来自正弦级数中 x 的更高次幂会对 T_1, T_3 及 T_5 项作出进一步的贡献. 但是这些贡献相对地小, 特别对前面的 T_k 项. 例如, x^5 项对 T_1 次的改动少于 1%, 而 x^7 项对它的改动少于 0.01%. 相反 x^5 对 T_3 次的改动约为 6%, 尽管 x^7 的贡献只有 0.02% 的额外量. 这就提示了, 截断我们的展开式将给我们对最小二乘三次式的一个好的逼近. 据此对我们的逼近取

$$\sin x\approx\frac{169}{192}T_1-\frac{5}{128}T_3\approx 0.9974x-0.1562x^3.$$

估计这逼近的精度时要注意我们曾作了二个"截断误差", 第一个是在 $\sin x$ 的幂级数中只取 3 项而第二个是丢弃了 T_5. 两者都影响第四位小数. 自然, 也可以用更大的精度, 假如我们寻求次数更高的最小二乘多项式, 然而即使是我们已有的这个逼近就有可与我们以它开始的 5 次 Taylor 多项式

相比美的精度.我们现在的 3 次多项式的误差与舍弃 x^5 项后所得到的 Taylor 三次多项式的误差之比较表示在图 21.8 上. Taylor 三次多项式在靠近零处较好,但是(几乎)最小二乘多项式的几乎等误差性质是明显的因而应与 $T_5(x)$ 作比较.

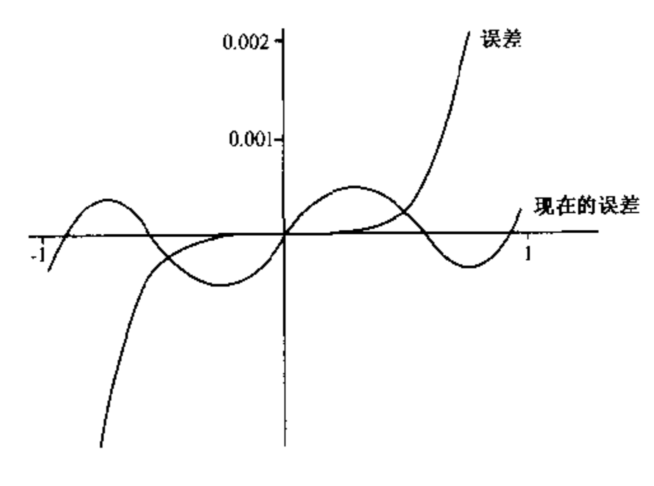


图 21.8

21.51　证明对小于 N 的 m 及 n 有

$$\sum_{i=0}^{N-1} T_m(x_i)T_n(x_i) = \begin{cases} 0, & m \neq n, \\ N/2, & m = n \neq 0, \\ N, & m = n = 0. \end{cases}$$

其中 $x_i = \cos A_i = \cos[(2i+1)\pi/2N]$, $i = 0, 1, \cdots, N-1$.

证　由 Chebyshev 多项式的三角定义,我们直接得到

$$\sum_{i=0}^{N-1} T_m(x_i)T_n(x_i) = \sum_{i=0}^{N-1} \cos mA_i \cos nA_i$$
$$= \frac{1}{2}\sum_{i=0}^{N-1}[\cos(m+n)A_i + \cos(m-n)A_i].$$

由于 $\cos a_i = \left(\frac{1}{2}\sin\frac{1}{2}a\right)\left[\Delta\sin a\left(i-\frac{1}{2}\right)\right]$故二个余弦和可以被缩短.然而,更为简单的是,注意到角 A_i 为 0 与 π 之间等间距分布的.除了当 $m+n$ 或 $m-n$ 为零之外每一个和均由于对称性而为零.这已经证明了当 $m \neq n$ 时的结果,若 $m = n \neq 0$,第二项和的贡献为 $\frac{N}{2}$,而若 $m = n = 0$ 时二个和总起来贡献为 N.应该注意到在求和号下的 Chebyshev 多项式与求积号下的一样是正交的.这常常是一种实质性的好处,因为对于复杂函数来说求和远比求积容易得多,特别当因子 $1/\sqrt{1-x^2}$ * 出现在积分中而不在求和中时.

21.52　选择什么样的系数 a_k 将会使

$$\sum_{x_i}[y(x_i) - a_iT_0(x_i) - \cdots - a_mT_m(x_i)]^2$$

极小化,其中 x_i 为上题中的自变量值?

解　以适当的标识符,直接由题 21.7 及 21.8 可知由

$$a_k = \frac{\sum y(x_i)T_k(x_i)}{\sum_i [T_n(x_i)]^2}$$

所决定的正交投影 $p = a_0T_0 + \cdots + a_mT_m$ 提供最小值.利用题 21.51 系数为

$$a_0 = \frac{1}{N}\sum_i y(x_i), \qquad a_k = \frac{2}{N}\sum_i y(x_i)T_k(x_i), \qquad k = 1, \cdots, m.$$

当 $m = N-1$ 时我们有关于 N 点 $(x_i, y(x_i))$ 的配置多项式,并且最小和为零.

* 译注:原文为 $\sqrt{1-x^2}$.

21.53　用题 21.52 的方法找出在区间(0,1)上对 $y=t^2$ 的最小二乘直线.

解　☞ 我们已经找到一根极小化题 21.46 中积分的直线.为了极小化题 21.52 的和,选 $t=(x+1)/2$ 如前.假设我们只用二个点,于是 $N=2$.这二点必定是 $x_0=\cos\pi/4=1/\sqrt{2}$ 及 $x_1=\cos 3\pi/4=-1/\sqrt{2}$.于是,

$$a_0=\frac{1}{2}\left[\frac{1}{8}(3+2\sqrt{2})+\frac{1}{8}(3-2\sqrt{2})\right]=\frac{3}{8},$$

$$a_1=\frac{1}{8}(3+2\sqrt{2})\left(\frac{1}{\sqrt{2}}\right)+\frac{1}{8}(3-2\sqrt{2})\left(-\frac{1}{\sqrt{2}}\right)=\frac{1}{2}.$$

而直线为 $p(x)=\frac{3}{8}T_0+\frac{1}{2}T_1=\frac{3}{8}+\frac{1}{2}x$ 所给出.这是与以前同样的一根直线并且当用一个更大的 N 时它还是会被再现,它可以简单地解释为 y 本身可以表示成 $y=a_0T_0+a_1T+a_2T_2$,并且因为 T_k 对积分及对和式都是正交的,故在二种意义下最小二乘直线也都可以由截断得到.(参看题 21.8 的最后一段.)

21.54　通过极小化题 21.52 的和找出在 $(-1,1)$ 上对 $y(x)=x^3$ 的最小二乘直线.

解　☞ 在这个问题中我们得到的直线将稍微依赖于所用的点数.首先取 $N=2$,这意味着像以前一样我们用 $x_0=-x_1=1/\sqrt{2}$.于是

$$a_0=\frac{1}{2}(x_0^3+x_1^3)=0,\qquad a_1=x_0^4+x_1^4=1/2.$$

选择 $N=3$ 我们得到 $x_0=\sqrt{3}/2$, $x_1=0$ 及 $x_2=-\sqrt{3}/2$.这就使得

$$a_0=\frac{1}{3}(x_0^3+x_1^3+x_2^3)=0,\qquad a_1=\frac{2}{3}(x_0^4+x_1^4+x_2^4)=3/4.$$

取 N 点的一般情况,我们有 $x_i=\cos A_i$ 并由于 A_i 在第一和第二象限中的对称性,

$$a_0=\frac{1}{N}\sum_{i=0}^{N-1}\cos^3A_i=0,$$

还有

$$a_1=\frac{2}{N}\sum_{i=0}^{N-1}\cos^4A_i=\frac{2}{N}\sum_{i=0}^{N-1}\left(\frac{3}{8}+\frac{1}{2}\cos 2A_i+\frac{1}{8}\cos 4A_i\right).$$

由于这些 A_i 是角 $\pi/2N, 3\pi/2N, \cdots(2N-1)\pi/2N$,其倍角为 $\pi/N, 3\pi/N, \cdots, (2N-1)\pi/N$ 因而这些倍角是围绕着整个圆周对称地分布的,$\cos 2A_i$ 的和因此为零.除 $N=2$ 外,$\cos 4A_i$ 的和也为零,所以对于 $N\neq 2$*, $a_i=\frac{3}{4}$.当 N 趋于无穷时,我们因此有对直线 $p(x)=3T_1/4=3x/4$ 的显见的收敛.

假如我们采用的是极小化积分的处理办法,则我们可得

$$a_0=\frac{1}{\pi}\int_{-1}^{1}\frac{x^3}{\sqrt{1-x^2}}dx=0,\qquad a_1=\frac{2}{\pi}\int_{-1}^{1}\frac{x^4}{\sqrt{1-x^2}}dx=\frac{3}{4}.$$

它导出的是同一根直线.

眼前的这个例子可以用来作为题 21.52 算法的一个更加基本的说明,但是只要注意到 $y=x^3=\frac{3}{4}T_1+\frac{1}{4}T_3$ 就更容易得到和理解这个结果,并且再一次求助于题 21.8 中的推论通过截断来得到 $3\pi/4$ 或 $3x/4$.截断过程当 $N=2$ 时不成立,因为从这时开始,T_0, T_1, T_2, T_3 不是正交的.(参看题 21.51.)

21.55　通过极小化题 21.52 中的和找出在 $(-1,1)$ 上对 $y=|x|$ 的最小二乘直线.

解　☞ 当 $N=2$ 时我们立刻得到 $a_0=1/\sqrt{2}$, $a_1=0$.取 $N=3$ 得结果 $a_0=1/\sqrt{3}$, $a_1=0$ 也是极容易的.对任意的 N

$$a_0=\frac{1}{N}\sum_{i=0}^{N-1}|\cos A_i|=\frac{2}{N}\sum_{i=0}^{I}\cos A_i,$$

其中 I 对奇数 N 为 $(N-3)/2$,对偶数 N 为 $(N-2)/2$.这个三角和数可以用缩短法或其他方法加以计算,其结果为

---

* 译注:原文为 $N=2$.

$$a_0 = \frac{\sin[\pi(I+1)/N]}{N\sin(\pi/2N)}.$$

对所有的 N，$a_1 = 0$，这是对称性的进一步的结果. 当 N 趋于无穷时有结果为

$$\lim a_0 = \lim \frac{1}{N\sin\pi/2N} = \frac{2}{\pi}.$$

当点用得越来越多时，就越趋向极限直线. 转向极小化积分的处理办法，我们当然预期这同一条直线. 计算的结果为

$$a_0 = \frac{1}{\pi}\int_{-1}^{1}\frac{|x|}{\sqrt{1-x^2}}dx = \frac{2}{\pi},$$

$$a_1 = \frac{2}{\pi}\int_{-1}^{1}\frac{x|x|}{\sqrt{1-x^2}}dx = 0.$$

因此正如我们所望. 极限直线为图 21.9 中的实线.

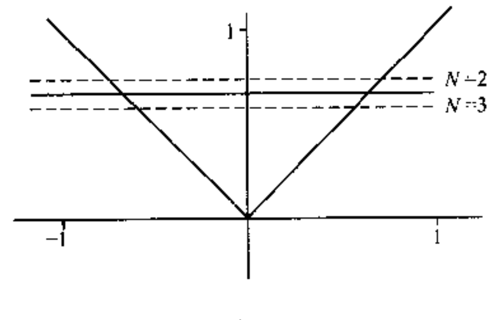


图 21.9

21.56 应用前面那些题中的方法于图 21.5 中实验得到的曲线.

解 对于这样一种具有未知解析特征的函数，我们的任何一种方法必须包含在某些点处的离散化. 我们为了使用 Simpson 法则已经选择了函数值的一个离散集合，因此至少在精神上保留了极小化一个积分的思想. 我们能用自变量的同一等距集合并极小化一个和式. 然而，以得到一个更接近等波纹误差的想法，现在我们选择自变量 $x_i = \cos A_i = 2t_i - 1$. 用前面用过的 11 个点，自变量 $x_i = \cos A_i = \cos[(2i+1)\pi/22]$ 与相应的 t_i 连同，从曲线上读出的 y_i 值如下：

x_i	0.99	0.91	0.75	0.54	0.28	0.00	−0.28	−0.54	−0.75	−0.91	−0.99
t_i	1.00	0.96	0.88	0.77	0.64	0.50	0.36	0.23	0.12	0.04	0.00
y_i	0.36	0.33	0.28	0.24	0.21	0.25	0.20	0.12	0.17	0.13	0.10

系数变成

$$a_0 = \frac{1}{11}\sum y_i \approx 0.22, \qquad a_1 = \frac{2}{11}\sum x_i y_i \approx 0.11,$$

使得直线 $p(x) = 0.22 + 0.11x = 0.22t + 0.11$，它与早先的结果几乎没有什么区别. 数据的不精确性没有造成额外的失真.

补 充 题

21.57 在一个 4 击洞处有不同障碍的高尔夫球被记录的平均得分如下：

障碍	6	8	10	12	14	16	18	20	22	24
平均分	4.6	4.8	4.6	4.9	5.0	5.4	5.1	5.5	5.6	6.0

找出关于这个数据的最小二乘直线.

21.58 用上题中的最小二乘直线来平滑被记录的数据.

21.59 估计每单位障碍平均得分的增长率.

21.60 对题 21.57 中的数据找出最小二乘抛物线.它是否与刚才所得到的直线有显著的不同?

21.61 当 x_i 及 y_i 都受到大致相同大小的误差时,曾经论证过,要极小化的是到一根直线垂直距离的平方和而不是纵向距离的平方和.证明要求极小化的是

$$S = \frac{1}{1-M^2}\sum_{i=0}^{N}(y_i - Mx_i - B)^2,$$

然后找出法方程并证明 M 由一个二次方程所决定.

21.62 应用上题的方法于题 21.57 的数据.新的直线是否与在那题中所得到的直线差别很大?

21.63 以题 21.1 的方法找出对于三个点(x_0,y_0),(x_1,y_1)及(x_2,y_2)的最小二乘直线.三个数 $y(x_i)-y_i$ 之符号的真实情况是什么?

21.64 证明对于数据

x_i	2.2	2.7	3.5	4.1
P_i	65	60	53	50

引进 $y=\log P$ 并计算关于数据对(x_i,y_i)的最小二乘直线将最后导出 $P=91.9x^{-0.43}$.

21.65 对数据

x_i	1	2	3	4
P_i	60	30	20	15

找出一个型如 $P=Ae^{Mx}$的函数.

21.66 仿照题 21.12 及 21.13 的过程证明用于 7 点的最小二乘抛物线导出平滑公式

$$y(x_k) \approx y_k - \frac{1}{21}(9\delta^4 y_k + 2\delta^6 y_k).$$

21.67 应用上面的公式对表 21.1 中央的 4 点 y_i 值进行平滑.与正确的方根值进行比较,并指明这个公式所产生的结果是否较 5 点公式的为好.

21.68 使用 7 点抛物线公式来导出近似微分公式

$$y'(x_k) \approx \frac{1}{28h}(-3y_{k-3} - 2y_{k-2} - y_{k-1} + y_{k+1} + 2y_{k+2} + 3y_{k+3}).$$

21.69 从来自表 21.1 的 y_i 值,应用上面公式对 $x=4,5,6$ 及 7 估计 $y'(x)$.这些结果与那些由 5 点抛物线所得到的相比情况如何?(参看题 21.19)

21.70 下面列出的是 $y(x)=x^2$ 带有附加从 -0.10 到 0.10 的随机误差的值.(误差是通过从一副去掉人头牌后的普通牌垛中抽出而得的,黑花代表加红花代表减.)准确值 T_i 也列在内.

x_i	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
y_i	0.98	1.23	1.40	1.72	1.86	2.17	2.55	2.82	3.28	3.54	3.92
T_i	1.00	1.21	1.44	1.69	1.96	2.25	2.56	2.89	3.24	3.61	4.00

应用题 21.13 与 21.15 的平滑公式,比较原始值与平滑值的标准差.

21.71 应用题 21.18 的微分公式于中央的 7 个自变量值.同时应用由 Stirling 多项式所得的公式(参看题 21.19).哪一个产生对 $y'(x)=2x$ 的更好逼近? 注意在本例中"真正"函数实际是抛物线,所以除了引入的随机误差外,我们会有精确结果.最小二乘抛物线是否已弥散误差到任何程度和已产生关于真正 $y'(x)$的信息?

21.72 什么是关于题 21.70 的数据的最小二乘抛物线? 将它与 $y(x)=x^2$ 比较.

21.73 利用题 21.20 的公式来估计在题 21.70 中所给出的数据表端点附近的 $y'(x)$.

21.74 由你计算所得的 $y'(x)$值来估计 $y''(x)$.

21.75 下面列出的是 $\sin x$ 带有附加从 -0.10 到 0.10 的随机误差的值.找出最小二乘抛物线并用它计算平

滑值.同时应用题 21.13 的方法来平滑数据,该方法是对每一点用了一个不同的最小二乘抛物线.哪一种方法更好?

x	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6
$\sin x$	−0.09	0.13	0.44	0.57	0.64	0.82	9.7	0.98	1.04

21.76 一个简单而又古老的平滑化方法,它依然是有用的,那就是移动平均法(moving averages).在这个方法中每个 y_i 值均被它和它附近点的平均值所代替.例如,假如在两侧各用二个邻近点,公式就是

$$p_i = \frac{1}{5}(y_{i-2} + y_{i-1} + y_i - y_{i+1} + y_{i+2}),$$

其中 p_i 为 y_i 的平滑过的替代值.应用它于上题的数据.设计一种平滑端点值的方法,对这种值来说有一侧缺少二个可用的邻点值.

21.77 对题 21.75 的数据应用移动平均法,在每侧只用一个邻点.关于内变量的公式将是

$$p_i = \frac{1}{3}(y_{i-1} + y_i + y_{i+1}),$$

设计一个对端点值进行平滑化的公式.

21.78 应用上题的公式对下面 $y(x) = x^3$ 的值.得到的 p_i 值被列出.

x_i	0	1	2	3	4	5	6	7
$y = x_i^3$	0	1	8	27	64	125	216	343
p_i		3	12	33	72	135	228	

证明这些 p_i 值属于一个不同的三次函数.应用移动平均公式于 p_i 值来得到第二代平滑值.假定在二端 y_i 值的供应可以无限制地扩张.你能说出在计算一代又一代平滑值时会发生什么情况?

21.79 应用移动平均法来平滑下面给出的振荡数据.

x_i	0	1	2	3	4	5	6	7	8
y_i	0	1	0	−1	0	1	0	−1	0

如果无止境地计算一代又一代平滑值时会发生什么情况?容易发现过多的平滑化会完全改变数据表的特征.

21.80 利用正交多项式去寻找在题 21.2 中获得的同一最小二乘直线.

21.81 利用正交多项式去寻找在题 21.10 中获得的同一最小二乘抛物线.

21.82 利用正交多项式去寻找一个用于题 21.14 中平方根数据的 4 次最小二乘多项式.利用这单个多项式去平滑数据,计算被光滑值的标准差.与那些在题 21.17 中所给出的进行比较.

21.83 下面是 e^x 带有附加从 −0.10 到 0.10 的随机误差的值.利用正交多项式寻找最小二乘三次式,这个三次式有多精确?

x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.10
y	0.92	1.15	1.22	1.44	1.44	1.66	1.79	1.98	2.32	2.51	2.81

21.84 下面是 Bessel 函数 $J_0(x)$ 带有附加从 −0.10 到 0.10 的随机误差的数值.利用正交多项式寻找一个最小二乘逼近.选择你认为恰当的次数.然后将数据平滑与同时提供的准确结果进行比较.

x	0	1	2	3	4	5	6	7	8	9	10
$y(x)$	0.994	0.761	0.225	−0.253	−0.400	0.170	0.161	0.301	0.177	−0.94	−0.240
准确值	1.00	0.765	0.224	−0.260	−0.397	0.178	0.151	0.300	0.172	−0.090	−0.246

21.85 找出关于 $y(x)=x^2$ 在区间$(-1,1)$的最小二乘直线.

21.86 找出关于 $y(x)=x^3$ 在区间$(-1,1)$上的最小二乘直线.

21.87 找出关于 $y(x)=x^3$ 在区间$(-1,1)$上的最小二乘抛物线.

21.88 用 Simpson 法则对积分进行估算,近似地找出关于图 21.10 中函数的最小二乘抛物线.这根曲线应该想象成是一个实验结果,理论断定它应当是一个抛物线.

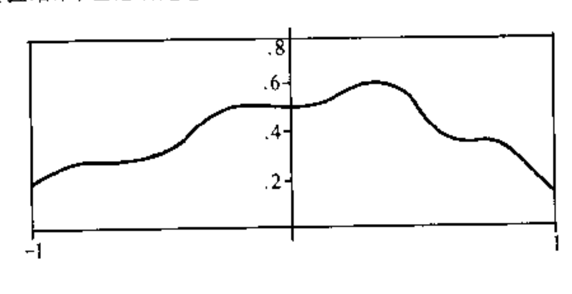


图 21.10

21.89 直接进行系数积分,证明关于 $\arcsin x$ 的 Chebyshev 级数为

$$\arcsin x=\frac{4}{\pi}\left(T_1+\frac{1}{9}T_3+\frac{1}{25}T_5+\frac{1}{49}T_7+\cdots\right),$$

在 T_3 后加以截断来得到该函数的最小二乘三次式.计算这个三次式的实际误差并与略去的第一项(T_5 项)进行比较.注意该误差的(几乎)等波纹性态.

21.90 寻找关于 $y(x)=x^2$ 在区间$(-1,1)$上带权函数 $w(x)=1/\sqrt{1-x^2}$的最小二乘直线,将这根直线与在题 21.85 中获得的那根进行比较.那一根具有等波纹性?

21.91 寻找关于 $y(x)=x^3$ 在区间$(-1,1)$上带权函数 $w(x)=1/\sqrt{1-x^2}$的最小二乘抛物线.将它与在题 21.87 中获得的抛物线进行比较.

21.92 重新将 $y(x)=e^{-x}$以其直至x^7项的幂级数来表示.对于 x 靠近 1 处这个误差在第 5 个小数位上.将这个和重新安排为 Chebyshev 多项式.应该取多少项才不会严重地影响第四位小数?重新安排这个被截断的多项式为标准形式.(这是多项式缩短的另一例.)

21.93 证明对于 $y(x)=T_n(x)=\cos(n\arccos x)=\cos nA$ 可得出 $y'(x)=(n\sin nA)/(\sin A)$.然后证明$(1-x^2)y''-xy'+n^2y=0$,这是 Chebyshev 多项式的经典的微分方程.

21.94 证明 $S_n(x)=\sin(n\arccos x)$也满足题 21.93 的微分方程.

21.95 令 $U_n(x)=S_n(x)/\sqrt{1-x^2}$并证明递推公式 $U_{n+1}(x)=2xU_n(x)-U_{n-1}(x)$.

21.96 验证 $U_0(x)=0$, $U_1(x)=1$,然后应用递推公式验证 $U_2(x)=2x$, $U_3(x)=4x^2-1$, $U_4(x)=8x^3-4x$, $U_5(x)=16x^4-12x^2+1$, $U_6(x)=32x^5-32x^3+6x$, $U_7(x)=64x^6-80x^4+24x^2-1$.

21.97 证明 $T_{m+n}(x)+T_{m-n}(x)=2T_m(x)T_n(x)$,然后令 $m=n$ 得到

$$T_{2n}(x)=2T_n^2(x)-1.$$

21.98 利用题 21.97 的结果来求得 T_8, T_{16}及 T_{32}.

21.99 证明$\frac{1}{n}T'_n=2T_{n-1}+\frac{1}{n-2}T'_{n-2}$,然后推出

$$T'_{2n+1}=2(2n+1)(T_{2n}+T_{2n-2}+\cdots+T_2)+1,$$

$$T'_{2n}=2(2n)(T_{2n-1}+T_{2n-3}+\cdots+T_1).$$

21.100 证明 $T_{2n+1}=x(2T_{2n}-2T_{2n-2}+2T_{2n-4}+\cdots\pm T_0)$.

21.101 通过重新安排成 Chebyshev 多项式将结果 $\ln(1+x)\approx x-\frac{1}{2}x^2+\frac{1}{3}x^3-\frac{1}{4}x^4+\frac{1}{5}x^5$ 减缩,然后仅保留二次项.证明最后结果 $\ln(1+x)\approx\frac{1}{32}+\frac{11}{8}x-\frac{3}{4}x^2$ 与原始逼近的 4 次部分具有大致相同的精

度.

21.102 将多项式 $y(x)=1+x+\frac{1}{2}x^2+\frac{1}{4}x^3+\frac{1}{24}x^4$ 减缩，首先将它表示为 Chebyshev 多项式的一个组合，然后在两项后加以截断.将这个结果与 $1+x+\frac{1}{2}x^2$ 进行比较，将两者考虑为 e^x 的逼近，哪一个是较好的逼近？在何种意义下？

21.103 证明变量变换 $x=2t-1$，它将区间转化成 t 的(0,1)区间，同时将 Chebyshev 多项式转化成下面的形式：

$$T_0^*(x)=1,\quad T_1^*(x)=2t-1,\quad T_2^*(x)=8t^2-8t+1,$$
$$T_3^*(x)=32t^3-48t^2+18t-1,\quad \text{等等}.$$

假如区间(0,1)被认为是更为方便.它可以用来替代经典多项式.还可证明递推公式

$$T_{n+1}^*(t)=(4t-2)T_n^*(t)-T_{n-1}^*(t).$$

21.104 证明 $\int T_0(x)\mathrm{d}x=T_1(x),\int T_1(x)\mathrm{d}x=\frac{1}{4}T_2(x)$，以及当 $n>1$ 时

$$\int T_n(x)\mathrm{d}x=\frac{1}{2}\left[\frac{T_{n+1}(x)}{n+1}-\frac{T_{n-1}(x)}{n-1}\right].$$

21.105 证明在题 21.53 中取 $N=2$ 时所得到的直线对任意的 N 也会出现相同的.

21.106 利用题 21.52 的方法来得到在(−1,1)上取 $N=3$ 时对 $y(x)=x^3$ 的一条最小二乘抛物线.证明对任意 N 以及用题 21.91 的极小化积分的方法所得结果相同.

21.107 寻找在区间(−1,1)上并且对任意 N，$y(x)=|x|$ 的最小二乘抛物线，同时证明当 $N\to\infty$ 时该抛物线趋于极小积分抛物线.

21.108 应用题 21.52 的方法于图 21.10 的实验数据.用这结果计算 $x=-1(0.2)1$ 处 $y(x)$ 的平滑值.

21.109 通过拟合一个最小二乘 5 次多项式来平滑下面的实验数据.

t	0	5	10	15	20	25	30	35	40	45	50
y	0	0.127	0.216	0.286	0.344	0.387	0.415	0.437	0.451	0.460	0.466

21.110 下面的表给出在一门考试中成绩为 x 的学生数 y.用这些结果作为一个标准模式，将 y 数用下面的平滑化公式

$$p=\frac{1}{35}(-3y_0+12y_1+17y_2+12y_3-3y_4)$$

平滑二次，假设对没有列出的 x 值 $y=0$.

x	100	95	90	85	80	75	70	65	60	55	50
y	0	13	69	147	208	195	195	126	130	118	121

x	45	40	35	30	25	20	15	10	5	0
y	85	95	75	54	42	30	34	10	8	1

21.111 对下面的数据找出最小二乘二次多项式，然后获得平滑过的值.

x	0.78	1.56	2.34	3.12	3.81
y	2.50	1.20	1.12	2.25	4.28

第二十二章 极小化极大多项式逼近

离散数据

通过多项式的极小化极大逼近的基本思想可以对离散数据表 x_i, y_i 的情况加以说明,其中 $i=1,\cdots,N$. 令 $p(x)$为一个不大于 n 次的多项式,并且令它与我们数据点之间的偏离量为 $h_i=p(x_i)-y_i$. 令 H 为这些"误差"中最大的一个:极小化极大多项式是指那个特殊的 $p(x)$,对它而言 H 为极小. 极小化极大逼近也被称为 Chebyshev 逼近. 主要的结果如下:

1. 存在性与惟一性　对任何给定的 n 值,极小化极大多项式的存在惟一性可以用下面描述的交换方法(exchange method)加以证明. 只对 $n=1$ 的情况提供细节.
2. 等误差性质是一个极小化极大多项式的标识性特征. 记该多项式为 $P(x)$,以及最大误差

$$E = \max|P(x_i) - y(x_i)|.$$

我们将证明 $P(x)$为仅有的一个多项式,对它而言 $P(x_i)-y(x_i)$以交替的符号,取极值 $\pm E$ 至少 $n+2$ 次.

3. 交换方法是一种通过它的等误差性质寻找 $P(x)$的算法. 选择某个含有 $n+2$ 个变量 x_i 的初始子集,对这些数据点找出一个等误差多项式. 若这个多项式在整个所选择的子集上的最大误差同时是它的全程极大值 H,那它就是 $P(x)$. 若不然则子集中的某个点被换成外部的一个点然后重复这一过程. 最后将被证明收敛于 $P(x)$.

连续数据

对于连续数据,习惯上几乎都是从回忆一个经典的被称作 Weierstrass 定理的分析定理开始,它陈述的是对于一个在区间(a,b)上的连续函数 $y(x)$,存在着一个多项式 $p(x)$使得

$$|p(x) - y(x)| \leqslant \varepsilon$$

在(a,b)中对任意的正数 ε 成立. 换言之,存在着一个多项式. 它以任何所要求的精度一致逼近 $y(x)$. 我们用 Bernstein 多项式来证明这个定理,它的形式为

$$B_n(x) = \sum_{k=0}^{n} p_{nk} y\left(\frac{k}{n}\right),$$

其中 $y(x)$是一个给定的函数而

$$p_{nk} = \binom{n}{k} x^k (1-x)^{n-k}.$$

我们对 Weierstrass 定理的证明包括证明 $\lim B_n(x)=y(x)$当 n 趋于无穷时一致地成立. Bernstein 多项式的收敛速度往往是令人失望的. 在实践中精确的一致逼近通常是由极小化极大方法获得.

极小化极大方法的基本事实或多或少与离散情况的那些相平行.

1. 对 $y(x)$的极小化极大逼近,是在所有不大于 n 次的多项式中,对给定的区间(a,b)极小化 $\max|p(x)-y(x)|$.
2. 它存在而且惟一.
3. 它具有等误差性质,是这类多项式中仅有的,对它而言 $p(x)-y(x)$在(a,b)中的 $n+2$ 个或更多的变量处取符号交错、大小为 E 的极值. 因此极小化极大多项式可以用它的等误差性加以识别. 在简单的例子中它可以完全地被显示出来. 当 $y''(x)>0$ 时极小化极大直线为其一例. 这里

$$P(x) = Mx + B,$$

取 $M=\frac{y(b)-y(a)}{b-a}$, $B=\frac{y(a)+y(x_2)}{2}-\frac{(a+x_2)[y(b)-y(a)]}{2(b-a)}$.

而 x_2 由

$$y'(x_2)=\frac{y(b)+y(a)}{b-a}$$

所确定.三个极值点为 a,x_2 及 b.然而,通常精确结果不是轻易能达到的,因而必须使用交换法来得到一个多项式,它接近于等误差性态.

4. Chebyshev 多项式级数,当它被截断时,时常产生具有几乎等误差性态的逼近.因此此类逼近为几乎极小化极大的.若它们本身不是完全适合的,则它们可以用作交换方法的初始输入,因此可望它比从一个更加任意的出发要收敛得快.

无穷范数

本章的根本主题是极小化范数

$$\|y-p\|_\infty,$$

其中 y 表示给定的数据而 p 表示逼近多项式.

题 解

离散数据,极小化极大直线

22.1 证明对于具有不同自变量 x_i 的任何三点 (x_i, Y_i) 确存在一条直线它对所有三个点的偏离值大小都相同而符号是交错的.这就是等误差直线或是 Chebyshev 直线.

证 ☞ 令 $y(x)=Mx+B$ 表示一条任意的直线并令 $h_i=y(x_i)-Y_i=y_i-Y_i$ 为在三个数据点处的"误差".一个易行的计算表明,因为 $y_i=Mx_i+B_i$,故对任何一条直线都有

$$(x_3-x_2)y_1-(x_3-x_1)y_2+(x_2-x_1)y_3=0.$$

定义 $\beta_1=x_3-x_2$,$\beta_2=x_3-x_1$,$\beta_3=x_2-x_1$,上面的方程变成

$$\beta_1y_1-\beta_2y_2+\beta_3y_3=0.$$

我们可以把它取成 $x_1<x_2<x_3$.所以三个 β 均为正数.我们要证明有一条直线,对它而言

$$h_1=h,\quad h_2=-h,\quad h_3=h,$$

使三个误差为大小相同但是符号交错.(这就是那条将被命名为"等误差"的直线.)现在,若一条具有这种性质的直线确实存在,那么

$$y_1=Y_1+h,\quad y_2=Y_2-h,\quad y_3=Y_3+h.$$

将它代入上式得

$$\beta_1(Y_1+h)-\beta_2(Y_2-h)+\beta_3(Y_3+h)=0,$$

就 h 解出

$$h=-\frac{\beta_1Y_1-\beta_2Y_2+\beta_3Y_3}{\beta_1+\beta_2+\beta_3}.$$

这已经证明了最多只有一条等误差直线存在,对于刚才计算所得的 h 值它必定通过三个点:(x_1, Y_1+h),(x_2, Y_2-h),(x_3, Y_3+h).虽然在正常情况下人们要求一条直线只通过二个指定的点,易见在眼前的这种特殊情况下,这三个点都要落在同一条直线上.P_1P_2 和 P_2P_3(其中 P_1,P_2,P_3 为从左到右的三点)的斜率为

$$\frac{Y_2-Y_1-2h}{x_2-x_1} \quad 及 \quad \frac{Y_3-Y_2+2h}{x_3-x_2},$$

利用我们早些时的方程易证它们是相等的.所以精确地有一根等误差直线,或者 Chebyshev 直线.

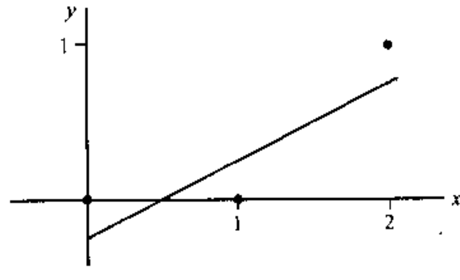


图 22.1

22.2 对数据点(0,0),(1,0)及(2,1)找出等误差直线.

解 ☞ 首先我们找出 $\beta_1=2-1=1,\beta_2=2-0=2,\beta_3=1-0=1$. 然后计算

$$h=-\frac{(1)(0)-(2)(0)+(1)(1)}{1+2+1}=-\frac{1}{4},$$

法直线通过 $\left(0,-\frac{1}{4}\right)\left(1,\frac{1}{4}\right)$ 及 $\left(2,\frac{3}{4}\right)$, 于是具有方程 $y(x)=\frac{1}{2}x-\frac{1}{4}$. 直线与点画在图 22.1 上.

22.3 证明对这三个点 (x_i, Y_i) 而言等误差直线也是极小化极大直线.

证 ☞ 等误差直线的误差为 $h,-h,h$. 令 h_1,h_2,h_3 为对任何其他直线的误差. 同时令 H 为 $|h_1|,|h_2|,|h_3|$ 三者中的最大者. 然后利用我们前面的公式

$$h=-\frac{\beta_1Y_1-\beta_2Y_2+\beta_3Y_3}{\beta_1+\beta_2+\beta_3}=-\frac{\beta_1(y_1-h_1)-\beta_2(y_2-h_2)+\beta_3(y_3-h_3)}{\beta_1+\beta_2+\beta_3},$$

其中, y_1,y_2,y_3 此时是归属于"任何其他直线". 将它重新整理为

$$h=-\frac{(\beta_1y_1-\beta_2y_2+\beta_3y_3)-(\beta_1h_1-\beta_2h_2+\beta_3h_3)}{\beta_1+\beta_2+\beta_3},$$

第一项为零我们得到等误差直线的 h 与另一条直线的 h_1,h_2,h_3 之间的关系,

$$h=\frac{\beta_1h_1-\beta_2h_2+\beta_3h_3}{\beta_1+\beta_2+\beta_3}.$$

由于诸 β 均为正的, 若我们将 h_1,h_2,h_3 分别换成 $H,-H$ 及 H, 则上式右端肯定会增加. 因此 $|h|\approx H$, 并且大小为 $|h|$ 的 Chebyshev 直线的最大误差从而不会大于其他任何直线的最大误差.

22.4 证明没有别的直线可能有与 Chebyshev 直线相同的最大误差, 所以极小化极大的直线是惟一的.

证 ☞ 假设在我们最后的结果中等式成立, $|h|=H$. 这就意味着以产生这个结果的 $H,-H,H$ 代入后并不会实际地增加 $\beta_1h_1-\beta_2h_2+\beta_3h_3$ 的大小. 但是这一点仅当 h_1,h_2,h_3 它们本身大小都相等并且有交错的符号时才会成立, 而这些特性带给我们的三点是 Chebyshev 直线所通过的三点. 肯定不会有二条直线通过这三个点. 这就证明了等式 $|h|=H$ 等同于 Chebyshev 直线. 如今我们已证明关于三点的等误差直线与极小化极大的直线是同一条.

22.5 通过对下面数据的应用来说明交换法(exchange method).

x_i	0	1	2	6	7
Y_i	0	0	1	2	3

解 ☞ 我们将简短地证明对 N 个点存在着惟一的一条极小化极大直线. 证明用了交换法, 它也是一个对计算这条直线而言为极好的算法, 基于这一点将首先说明这个方法. 它包含了 4 个步骤:

第 1 步 选择数据点的任意三个(三个数据点的这种集合将被称作一个三元组(triple). 在这一步上简单地选一个初始三元组. 它将在第 4 步中被改变.)

第 2 步 对这个三元组找出 Chebyshev 直线. 关于该直线的 h 值当然在这个过程中加以计算.

第 3 步 对刚找到的 Chebyshev 直线计算所有数据点的误差. 记这些 h_i 值中(取绝对值)最大的为 H. 若 $|h|=H$ 则寻查结束. 关于当前的这个三元组的 Chebyshev 直线就是对 N 个点的整个集合的极小化极大的直线. (我们将简短地证明这一点.) 若 $|h|<H$ 则前进到第 4 步.

第 4 步 这是交换的一步. 选择一个新的三元组如下. 在老的三元组上加上一个数据点, 在该点处出现大小为 H 的最大的误差. 然后在先前的点中去除一点, 按照保留的三点其误差有交错的符号的方法. (短暂的实践将表明这样做总是可能的.) 以新的三元组返回到第 2 步和第 3 步.

为了说明这一点我们选初始三元组为

$$(0,0)\qquad(1,0)\qquad(2,1)$$

由头三个点所组成. 这是题 22.2 的三元组, 对它而言我们已经找到 Chebyshev 直线为 $y=\frac{1}{2}x-\frac{1}{4}$ 取 $h=-\frac{1}{4}$. 它完成了第一步和第二步. 前进到第 3 步我们获得在所有 5 个数据点上的误差为 $-\frac{1}{4}$,

$\frac{1}{4}, -\frac{1}{4}, \frac{3}{4}, \frac{1}{4}$. 它使得 $H=h_4=\frac{3}{4}$. 这条 Chebyshev 直线在它已有的三元组上是一个等误差直线，但它偏离第 4 个数据点比较多.（参看图 22.2 中的虚线.）

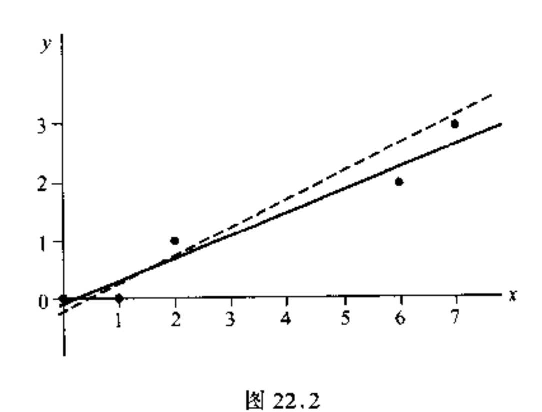


图 22.2

因此转向第 4 步，现在我们将第 4 点包括进来而去掉第一点来获得新三元组

$$(1,0)\qquad(2,1)\qquad(6,2)$$

在它上面老 Chebyshev 直线的误差确实有所要求的交错符号 $\left(\frac{1}{4}, -\frac{1}{4}, \frac{3}{4}\right)$. 以这个三元组我们返回到第 2 步并找到一条新的 Chebyshev 直线. 从计算

$$\beta_1 = 6-2=4,\qquad \beta_2=6-1=5,\qquad \beta_3=2-1=1,$$

$$h=-\frac{(4)(0)-(5)(1)+(1)(2)}{4+5+1}=\frac{3}{10}$$

开始，于是直线必须通过 $\left(1,\frac{3}{10}\right)$，$\left(2,\frac{7}{10}\right)$ 及 $\left(6,\frac{23}{10}\right)$ 这三点. 找到的这条直线为 $y=\frac{2}{5}x-\frac{1}{10}$. 重复第 3 步我们得到 $-\frac{1}{10}, \frac{3}{10}, -\frac{3}{10}, \frac{3}{10}, -\frac{3}{10}$ 这 5 个误差；并由于 $H=\frac{3}{10}=|h|$，作业便完成了.

对于这个新的三元组的 Chebyshev 直线是关于整个点集的极小化极大直线. 它的最大误差是 $\frac{3}{10}$. 新的直线是图 22.2 中所示的实线. 注意我们新直线的 $|h|$ 值 $\left(\frac{3}{10}\right)$ 大于第一条直线的 $\left(\frac{1}{4}\right)$. 但是在整个点集上最大误差从 $\frac{3}{4}$ 减至 $\frac{3}{10}$，因而它就是极小化极大误差. 将对一般情况证明这一点.

22.6 证明交换法第 3 步的条件 $|h|=H$ 最终会被满足，所以方法将停止.（设想我们永远可以作交换.）

证 ☞ 回忆在任何特定的交换后老的 Chebyshev 直线在新三元组上的误差其大小为 $|h|$, $|h|$, H. 同时回顾 $|h|<H$（或者是我们已经停止）以及三个误差有不同的符号. 于是对这个新的三元组找到了 Chebyshev 直线. 称它的在这个新的三元组上的误差为 $h^*, -h^*, h^*$. 返回到题 22.3 中关于 h 的公式，用老 Chebyshev 直线扮演“任何其他直线”的角色，我们有

$$h^*=\frac{\beta_1h_1-\beta_2h_2+\beta_3h_3}{\beta_1+\beta_2+\beta_3},$$

其中 h_1, h_2, h_3 是具有交错符号的 h, h, H. 因为符号的交错，故在分子中的所有三项有相同的符号，因而

$$|h^*|=\frac{\beta_1|h|+\beta_2|h|+\beta_3H}{\beta_1+\beta_2+\beta_3},$$

这里我们假设误差 H 正是在所指定的第三点上（它走向哪个位置没有实际差别.）在任何情况下因为 $H>|h|$ 故 $|h^*|>|h|$. 新的 Chebyshev 直线在它的三元组上有比老的在它的组上更大的误差，这个结果如今提供了极好的服务. 若它的出现是一个惊讶的话，那就这样来看待它. 老的直线在它自己的三元组上给出了极好的服务 $\left(在我们的例中\ h=\frac{1}{4}\right)$ 但是在其他地方服务得不好 $\left(H=\frac{3}{4}\right)$. 新的直线在它自己的三元组上给出了好的服务 $\left(h=\frac{3}{10}\right)$. 并且在其他点上也有同样好的服务.

现在我们可以证明交换法必定有个终了时刻. 因为只有这么多个三元组. 而且没有一个三元组会被二次中选，因为正如刚才证明过的，h 值要持续增长. 在某个阶段上 $|h|=H$ 终将会被满足.

22.7　证明以交换法最后计算得的 Chebyshev 直线是对整个 N 点集合的极小化极大直线.

证 ☞ 令 h 为最后的 Chebyshev 直线在它自己的三元组上的等误差值. 于是在整个点集上的最大误差其大小为 $H=|h|$, 否则我们还需通过另一次交换进行到另一个三元组和另一条直线. 令 $h_1, h_2, \cdots, h_N$ 为对于任何一条其他直线的误差. 于是 $|h|<\max|h_i|$, 其中 h_i 限于最后一个三元组的三个点, 因为没有一条直线在它自己的三元组上充当一条 Chebyshev 直线. 但是这么一来对于不加限制的 h_i 而言肯定地也有 $|h|<\max|h_i|$, 因为包含 N 点中剩下的那些点只会使右侧甚至更大一些. 因此 $H=|h|<\max|h_i|$ 而最后的 Chebyshev 直线的最大误差是所有的最大误差中最小的. 总而言之, 关于 N 个点的集合上的极小化极大的直线是一条在适当选择的三元组上的等误差直线.

22.8　应用交换法来获得关于下面数据的极小化极大的直线:

x_i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y_i	0	1	1	2	1	3	2	2	3	5	3	4	5	4	5	6

x_i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Y_i	6	5	7	6	8	7	7	8	7	9	11	10	12	11	13

解 ☞ 可用的三元组的数目为 $C(31,3)=4495$, 所以要找到准确值人们可以将它与海底捞针相比拟. 然而, 交换法在无意义的三元组上花的时间十分地少. 从在 $x=(0,1,2)$ 处的非常差的三元组开始, 只需三次交换就产生极小化极大的直线 $y(x)=0.38x-0.29$, 它的系数被舍入成两位. 逐次的带有 h 与 H 值的三元组如下:

三元组在 $x=$	(0,1,2)	(0,1,24)	(1,24,30)	(9,24,30)
h	0.250	0.354	−1.759	−1.857
H	5.250	3.896	2.448	1.857

注意在本例中没有一次把不需要的点带入三元组中. 需要三个点, 三次交换就达到目的. 还要注意到像预告的那样 $|h|$ 稳步地增长. 这 31 个点, 极小化极大直线, 以及最终的三元组(虚的垂线表示等误差)画在图 22.3 上.

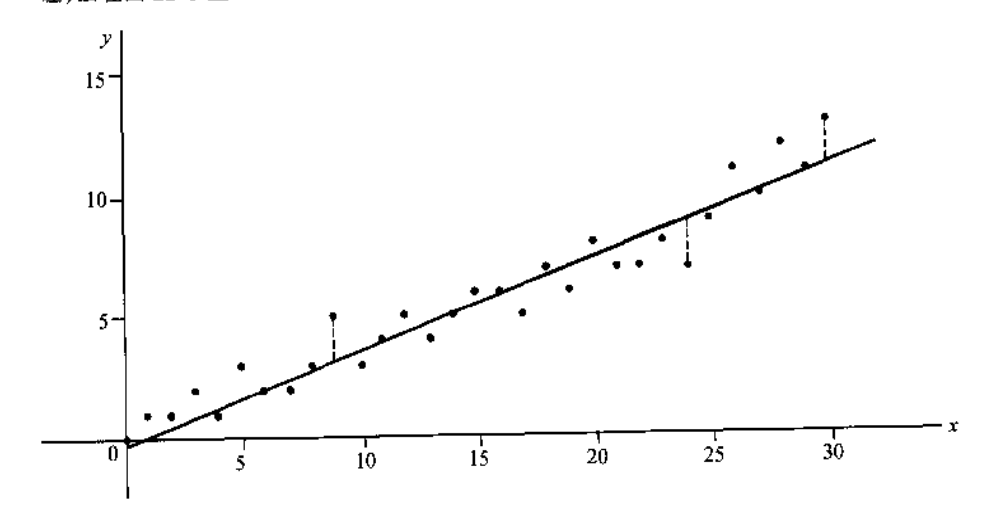


图 22.3

离散数据，极小化极大多项式

22.9　将交换法推广到对下面的数据寻找极小化极大抛物线.

x_i	-2	-1	0	1	2
y_i	2	1	0	1	2

解　当然数据是来自函数 $y=|x|$，但是这个简单的函数将作为例子，用来说明交换法的基本思想是如何从刚才处理过的直线问题推向去发现一个极小化极大多项式.这类多项式的存在性，惟一性以及等误差性质只是我们为极小化极大直线所作证明的推广，因此不在这里给出.算法现在由选一个"初始四元组"开始，并且我们将头四点取作 $x=-2,-1,0,1$，对这个四元组我们寻找一个等误差抛物线，譬如说它就是

$$p_1(x)=a+bx+cx^2,$$

这意味着我们要求交替地有 $p(x_i)-y_i=\pm h$，或是

$$a-2b+4c-2=h,$$
$$a-b+c-1=-h,$$
$$a-0=h,$$
$$a+b+c-1=-h,$$

解这四个方程，我们得到 $a=\frac{1}{4},b=0,c=\frac{1}{2},h=\frac{1}{4}$，因而 $p_1(x)=\frac{1}{4}+\frac{1}{2}x^2$.这就完成了等价的第1及第2步，然后我们转向第3步并计算我们的抛物线在所有5个数据点处的误差，它们是$\frac{1}{4}$，$-\frac{1}{4}$，$\frac{1}{4}$，$-\frac{1}{4}$，$\frac{1}{4}$.因而在整个集合上的最大误差$\left(H=\frac{1}{4}\right)$等于在我们四元组上的最大误差$\left(|h|=\frac{1}{4}\right)$.算法结束而我们的第一个抛物线就是极小化极大的抛物线.它如图22.4所示.

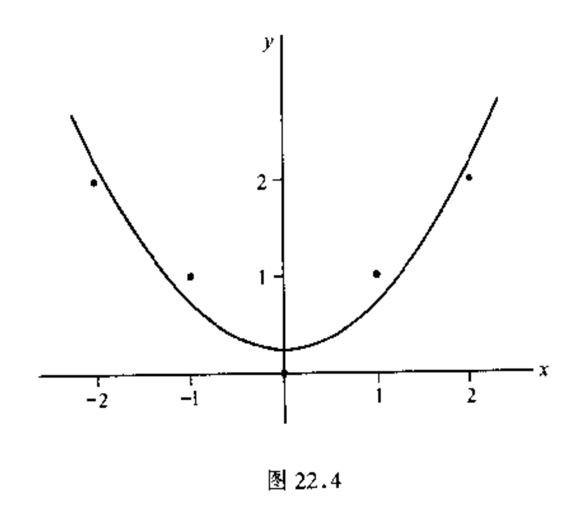


图 22.4

22.10　找出关于7个点 $y=|x|,x=-3(1)3$ 的极小-极大抛物线.

解　它在我们前面的数据表的两端各加上一个点.假设我们选择初始四元组如前.那么我们再一次获得上题中的等误差抛物线 $p_1(x)$，在新的数据点处的误差为$\frac{7}{4}$，所以现在 $H=\frac{7}{4}$而$|h|=\frac{1}{4}$.据此我们引入一个新点进入四元组并舍弃$x=-2$.在新的四元组上，老抛物线具有误差$-\frac{1}{4}$，$\frac{1}{4}$，$-\frac{1}{4}$，$\frac{7}{4}$它们确有交替的符号.作了这种交换后，一定会获得一个新的等误差抛物线

$$p_2(x)=a_2+b_2x+c_2x^2.$$

如在上题中那样做法我们立刻得到等误差 $h_2=-\frac{1}{3}$ 以及抛物线 $p_2(x)=\frac{1}{3}(1+x^2)$. 它在 7 个数据点上的误差为 $\frac{1}{3}$, $-\frac{1}{3}$, $\frac{1}{3}$, $-\frac{1}{3}$, $\frac{1}{3}$, $-\frac{1}{3}$, $\frac{1}{3}$, 因而 $H=|h|=\frac{1}{3}$ 算法终止. 该抛物线 $p_2(x)$ 是极小化极大抛物线. 所有误差都有一样的大小则是一个意外的收获. 这对极小化极大多项式而言并非一般性的特征, 正像刚解出的直线问题所展示的那样.

连续数据, Weierstrass 定理

22.11　证明 $\sum_{k=0}^{n} p_{nk}^{(x)}(k-nx)=0$ 其中 $p_{nk}^{(x)}=\binom{n}{k}x^k(1-x)^{n-k}$.

证 ☞ 关于整数 n 及 k 的二项式定理

$$(p+q)^n=\sum_{k=0}^{n}\binom{n}{k}p^kq^{n-k}$$

是 p 和 q 的一个恒等式. 对 p 微分得到

$$n(p+q)^{n-1}=\sum_{k=0}^{n}\binom{n}{k}kp^{k-1}q^{n-k},$$

以 p 乘它然后令 $p=x$, $q=1-x$, 它就变成 $nx=\sum_{k=0}^{n}kp_{nk}^{(x)}$. 用同样的 p 及 q 于二项式定理本身表明 $1=\sum p_{nk}^{(x)}$, 于是最终得到

$$\sum_{k=0}^{n}p_{nk}^{(x)}(k-nx)=nx-nx=0.$$

22.12　还证明 $\sum_{k=0}^{n}p_{nk}^{(x)}(k-nx)^2=nx(1-x)$.

证 ☞ 对 p 的二次微分得到

$$n(n-1)(p+q)^{n-2}=\sum_{k=0}^{n}\binom{n}{k}k(k-1)p^{k-2}q^{n-k},$$

以 p^2 乘它然后令 $p=x$, $q=1-x$, 它就变成

$$n(n-1)x^2=\sum_{k=0}^{n}k(k-1)p_{nk}^{(x)},$$

从它我们得到

$$\sum_{k=0}^{n}k^2p_{nk}^{x}=n(n-1)x^2+\sum_{k=0}^{n}kp_{nk}^{(x)}=n(n-1)x^2+nx.$$

最后我们计算

$$\sum_{k=0}^{n}p_{nk}^{(x)}(k-nx)^2=\sum k^2p_{nk}^{(x)}-2nx\sum kp_{nk}^{(x)}+n^2x^2\sum p_{nk}^{(x)}$$

$$=n(n-1)x^2+nx-2nx(nx)+n^2x^2=nx(1-x).$$

22.13　证明若 $d>0$ 且 $0\leqslant x\leqslant 1$, 则

$$\sum{}'p_{nk}^{(x)}\leqslant\frac{x(1-x)}{nd^2}$$

其中 $\sum'$ 是对满足 $|(k/n)-x|\geqslant d$ 的 k 求和. (这是有名的 Chebyshev 不等式的一个特殊情况.)

证 ☞ 将上题的和分成二个部分

$$nx(1-x)=\sum{}'p_{nk}^{(x)}(k-nx)^2+\sum{}''p_{nk}^{(x)}(k-nx)^2,$$

其中 $\sum''$ 包含了在 $\sum'$ 中未出现的 k. 因而

$$nx(1-x)\geqslant\sum{}'p_{nk}^{(x)}(k-nx)^2$$

$$\geqslant\sum{}'p_{nk}^{(x)}n^2d^2.$$

这里第一个不等式成为可能是由于 $\sum''$ 为非负的, 第二个则因为在 $\sum'$ 中我们有 $|k-nx|\geqslant nd$. 以 n^2d^2 通除上式, 我们便得到所要的结果.

22.14　导出对$\sum'$及$\sum''$的这些估计：

$$\sum{}'p_{nk}^{(x)} \leqslant \frac{1}{4nd^2}, \qquad \sum{}''p_{nk}^{(x)} \geqslant 1 - \frac{1}{4nd^2}.$$

解　☞ 函数 $x(1-x)$在 $x=\frac{1}{2}$处取它的最大值并因此当 $0\leqslant x\leqslant 1$ 时 $0\leqslant x(1-x)\leqslant\frac{1}{4}$. $\sum'$的结果是上题的立刻可得的结论.而因此$\sum''=1-\sum'\geqslant 1-(1/4nd^2)$.

22.15　证明若 $f(x)$当$0\leqslant x\leqslant 1$ 时为连续的，则当 n 趋向无穷时 $\lim\sum_{k=0}^{n}p_{nk}^{(x)}f(k/n)=f(x)$ 一致地成立.

证　☞ 通过展示一个一致收敛于 $f(x)$的多项式序列

$$B_n(x)=\sum_{k=0}^{n}p_{nk}^{(x)}f\left(\frac{k}{n}\right)$$

将证明 Weierstrass 定理.这些多项式被称为关于 $f(x)$的 Bernstein 多项式.证明从选择一个任意的正数 ϵ 开始，于是当$|x'-x|<d$ 时有

$$|f(x')-f(x)|<\frac{\epsilon}{2}$$

并且 d 由于$f(x)$的一致连续性而与 x 无关.然后以 M 表示$|f(x)|$的最大值，我们有

$$\begin{aligned}|B_n(x)-f(x)| &= \left|\sum{}'p_{nk}^{(x)}\left[f\left(\frac{k}{n}\right)-f(x)\right]\right|\\ &\leqslant\sum{}'p_{nk}^{(x)}\left|f\left(\frac{k}{n}\right)-f(x)\right|\\ &\quad+\sum{}''p_{nk}^{(x)}\left|f\left(\frac{k}{n}\right)-f(x)\right|\\ &\leqslant 2M\sum{}'p_{nk}^{(x)}+\frac{1}{2}\epsilon\sum{}''p_{nk}^{(x)},\end{aligned}$$

这里 k/n 在$\sum''$的部分扮演x'的角色，$\sum''$的定义保证了$|x'-x|<d$.于是对于充分大的 n 有

$$\begin{aligned}|B_n(x)-f(x)| &\leqslant\left(\frac{2M}{4nd^2}\right)+\frac{1}{2}\epsilon\\ &\leqslant\frac{1}{2}\epsilon+\frac{1}{2}\epsilon=\epsilon,\end{aligned}$$

这就是所要求的结果.对于非(0,1)的其他区间可以通过一个简单的变量变换予以调节.

22.16　证明在 $f(x)=x^2$ 的情况下，$B_n(x)=x^2+x(1-x)/n$，所以 Bernstein 多项式并不是对 $f(x)$的给定次数的最佳逼近.[肯定地说对 $f(x)=x^2$ 的最佳二次逼近就是 x^2 自己.]

证　☞ 由于在题 22.12* 中已获得和$\sum k^2p_{nk}^{(x)}$，故

$$\begin{aligned}B_n(x) &= \sum_{k=0}^{n}p_{nk}^{(x)}f\left(\frac{k}{n}\right)=\sum_{k=0}^{n}\frac{p_{nk}^{(x)}k^2}{n^2}=\frac{1}{n^2}[n(n-1)x^2+nx]\\ &= x^2+\frac{x(1-x)}{n}\end{aligned}$$

就是所要求的.当 n 趋向无穷时一致收敛是显而易见的，但是，显然 $B_n(x)$并不能完全复制 x^2.我们现在转向更好的一类一致逼近多项式.

连续数据，Chebyshev 理论

22.17　证明若 $y(x)$在 $a\leqslant x\leqslant b$ 中连续，则存在着一个不大于 n 次的多项式 $P(x)$使得 $\max|P(x)-y(x)|$在区间(a,b)上是一个极小.换言之，没有别的这种类型的多项式会产生一个更小的极大.

证　☞ 令 $p(x)=a_0+a_1x+\cdots+a_nx^n$ 为任何次数不大于 n 的多项式.此时

$$M(a)=\max|p(x)-y(x)|$$

---

* 译注：原文为 22.2.

依赖于所选定的多项式 $p(x)$，也就是说，它依赖于我们以 $\bar{a}$ 作为标记的系数集合 $(a_0, a_1, \cdots, a_n)$. 由于 $M(\bar{a})$ 是 $\bar{a}$ 的一个连续函数而且是非负的，所以它有一个最大的下界. 记这个界为 L. 必须要证明的是对某个特定的系数集合 A，($P(x)$ 的系数)，下界 L 是真正达到的，即 $M(A)=L$. 作为对比，函数 $f(t)=1/t$ 对正的 t 最大的下界为零，但是没有一个自变量值 t 使 $f(t)$ 真正达到这个界，t 的范围无界自然是允许这种情况出现的一个因素. 在我们的问题中系数集合 $\bar{a}$ 也有无限的范围，但是我们现在证明，还是有 $M(A)=L$. 作为开始，对 $i=0,1,\cdots,n$ 令 $a_i=Cb_i$ 使其满足 $\sum b_i^2=1$. 我们也可以记 $\bar{a}=C\bar{a}$. 考虑第二个函数

$$m(\bar{b}) = \max|b_0+b_1x+\cdots+b_nx^n|,$$

其中 max 是指像通常的在区间 (a,b) 上多项式的极大值. 这是一个在单位球 $\sum b_i^2=1$ 上的连续函数. 在这样一个集合上(闭且有界)一个连续函数的确能取到它的最小值. 记这个极小值为 μ. 清楚地 $\mu\geqslant0$. 但这个零值是不可能的，因为只有 $p(x)=0$ 才会产生这个极小值，而加在 b_i 上的条件临时将这个多项式排除在外了. 因此 $\mu>0$. 从而

$$\begin{aligned} m(\bar{a}) &= \max|a_0+a_1x+\cdots+a_nx^n| \\ &= \max|p(x)| = Cm(\bar{b}) \geqslant C\mu. \end{aligned}$$

现在回到 $M(\bar{a})=\max|p(x)-y(x)|$，并利用差的绝对值超过绝对值的差，我们得到

$$\begin{aligned} M(\bar{a}) &\geqslant m(\bar{a}) - \max|y(x)| \\ &\geqslant C\mu - \max|y(x)|. \end{aligned}$$

若我们选择 $C>(L+1+\max|y(x)|)/\mu=R$，则立得 $M(\bar{a})\geqslant L+1$. 回忆起 L 是 $M(\bar{a})$ 的最大的下界，我们看到 $M(a)$ 当 $C>R$ 时比 L 要大而它在约束 $C\leqslant R$ 之下的最大的下界将同样是这个 L. 但是这个约束等价于 $\sum a_i^2\leqslant R$，故如今又出现这样的情况：一个连续函数 $M(\bar{a})$ 在一个有界闭集上(一个实心的球体，或球). 在这样的一个集合上最大下界是真正可以取到的，譬如说在 $\bar{a}=A$ 处. 因此 $M(A)$ 就是 L，而 $p(x)$ 就是极小化极大多项式.

22.18 令 $P(x)$ 为在区间 (a,b) 上在所有次数不大于 n 的多项式中逼近 $y(x)$ 的一个极小化极大多项式. 令 $E=\max|y(x)-P(x)|$，并假设 $y(x)$ 本身不是一个次数不大于 n 的多项式，因此 $E>0$，说明至少有一个自变量，对它来说 $y(x)-P(x)=E$ 成立，并且对 $-E$ 类似地成立. [我们继续假设 $y(x)$ 连续.]

解 由于 $y(x)-P(x)$ 在 $a\leqslant x\leqslant b$ 中连续，它一定会在某处达到 $\pm E$ 之一. 我们要证明二者一定都会达到. 假设在 (a,b) 中任何地方都不等于 E. 此时

$$\max[y(x)-P(x)] = E-d,$$

其中 d 是正的. 于是有

$$-E \leqslant y(x) - P(x) \leqslant E - d$$

而这可以写成

$$-E+\frac{1}{2}d \leqslant y(x) - \left[P(x) - \frac{1}{2}d\right] \leqslant E - \frac{1}{2}d.$$

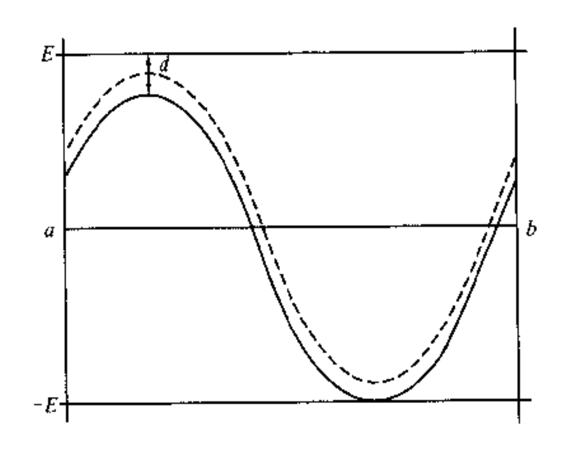


图 22.5

它断然地声称 $P(x)-\frac{1}{2}d$ 逼近 $y(x)$ 以一个 $E-\frac{1}{2}d$ 的最大误差.这与最初假设 $P(x)$ 本身是一个具有最大误差 E 的极小化极大多项式相矛盾.因此 $y-P(x)$ 必定在 (a,b) 中的某个地方等于 E.一个十分类似的证明来指出它必定也等于 $-E$.图 22.5 说明了这个证明的简单思路.对于极小化极大多项式而言其误差不可能有如实线所示的性态,因为将曲线升高 $\frac{1}{2}d$ 就会带来一个新的误差曲线(虚线所示)具有一个较小的最大绝对值 $E-\frac{1}{2}d$,而这是一个矛盾.

22.19 继续上题,证明对于 $n=1$,即以线性多项式逼近,必定存在一个第三点,在该点处极小化极大 $P(x)$ 之误差 $|y(x)-P(x)|$ 取它的最大值 E.

证 令 $y(x)-P(x)=E(x)$ 并将 (a,b) 分成子区间到足够小,使得对在任何子区间内的 x_1,x_2 都有

$$|E(x_1)-E(x_2)|\leqslant\frac{1}{2}E.$$

由于 $E(x)$ 在 $a\leqslant x\leqslant b$ 中连续,故上式肯定能成立.在一个子区间中,称之为 I_1,我们知道误差达到 E,譬如说在 $x=x_+$ 处.由此得出在整个子区间中

$$|E(x)-E(x_+)|=|E(x)-E|\leqslant\frac{1}{2}E,$$

使得 $E(x)\geqslant\frac{1}{2}E$.类似地,在一个子区间中,称之为 I_2,我们获得 $E(x_-)=-E$,从而 $|E(x)|\leqslant-\frac{1}{2}E$.因此这两个区间不可能是相邻的,于是我们可以在它们之间选择一点 u_1.假设 I_1 在 I_2 的左边(在相反的情况下讨论几乎是完全相同的)于是 u_1-x 与在所考虑的两个子区间的每一个中都与 $E(x)$ 有相同的符号.令 $R=\max|u_1-x|$ 其中 x 在 (a,b) 中.

现在假设不存在第三个点,在该点处误差为 $\pm E$.于是在所有的子区间中除了刚才所考虑过的二个区间外我们一定有

$$\max|E(x)|<E.$$

而且由于有有限多个子区间

$$\max[\max|E(x)|]=E^*<E,$$

自然地 $E^*\geqslant\frac{1}{2}E$ 因为这些子区间会扩展到 I_1 和 I_2 的端点,在那里 $|E(x)|\geqslant\frac{1}{2}E$.考虑 $P(x)$ 的下面的替代品,它还是一个线性多项式:

$$P^*(x)=P(x)+\epsilon(u_1-x).$$

若我们选 ϵ 充分小使得 $\epsilon R<E-E^*\leqslant\frac{1}{2}E$,则 $P^*(x)$ 变成一个比 $P(x)$ 更好的逼近.因为

$$|y(x)-P^*(x)|=|E(x)-\epsilon(u_1-x)|,$$

所以在 I_1 中误差是减少了但仍为正的,同时在 I_2 中误差增加了但保留为负的;在这二个子区间中误差的大小是减下来了.在别处,虽然误差的大小可能增长,它不可能超过 $E^*+\epsilon R<E$,因而 $P^*(x)$ 有一个比 $P(x)$ 更小的最大误差.这个矛盾说明一个具有误差 $\pm E$ 的第三点一定存在.图 22.6

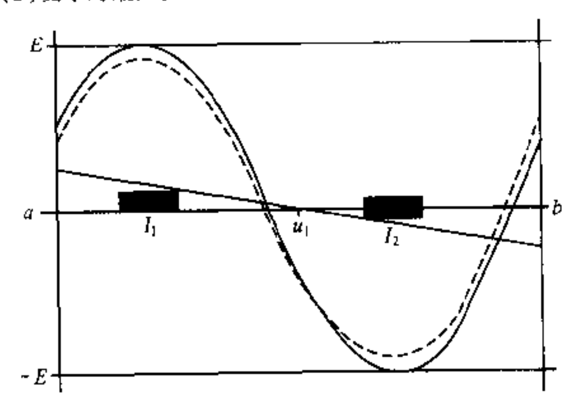


图 22.6

说明在这个证明背后的简单思路. 误差曲线 $E(x)$ 不可能有实的曲线那样的性态. (只有两个 $\pm E$ 点). 因为加上了线性校正项 $\epsilon(u_1 - x)$ 到 $P(x)$ 上就将误差减少了这个同样的量, 引出一个新的误差曲线(用虚线表示)具有较小的最大绝对值.

22.20 证明对上题中的 $P(x)$, 必定存在着三个点在这些点处其误差大小为 E 而符号交错地出现.

证 ☞ 上题的证明已经很充分了. 例如, 如果符号是 $+, +, -$, 那么就选 u_1 在相邻的 $+$ 和 $-$ 之间我们的 $P^*(x)$ 也就比 $P(x)$ 好. 完全相同的说明也适用于模式 $+, -, -$, 只有符号交错才能避免矛盾.

22.21 证明, 在一般情况下不大于 n 次的极小化极大的多项式肯定存在 $n+2$ 个具有交错符号的误差最大的点.

证 ☞ 它的证明可通过处理 $n=2$ 的情况加以说明. 令 $P(x)$ 为一个不大于 2 次的极小化极大的多项式, 由题 22.18 它肯定地至少有二个最大误差点. 题 22.19 与 22.20 中的论点现在除了以二次的 $P(x)$ 来代替线性的之外别无更改, 于是证明了一定存在第三个这种点而且符号一定是交错的, 譬如说 $+, -, +$ 刚好就是要定的. 现在假设没有第四处最大误差出现. 我们重复题 22.19 的论点, 在出现误差 $\pm E$ 的区间 I_1, I_2 及 I_3 之间选择二点 u_1 及 u_2, 并利用校正项 $\epsilon(u_1 - x)(u_2 - x)$, 它在符号上与这些子区间中的 $E(x)$ 一致. 不需要作其他的改变. 二次的 $P^*(x)$ 将有一个比 $P(x)$ 更小的最大误差, 而这个矛盾证明了第四个 $\pm E$ 点必定存在. 符号的交错由用在题 22.20 中的相同论点来完成, 而推广到更高的 n 值是完全类似的.

22.22 证明对每个 n 只有一个极小化极大的多项式.

证 ☞ 假设有二个 $P_1(x)$ 及 $P_2(x)$, 则

$$-E \leqslant y(x) - P_1(x) \leqslant E, \qquad -E \leqslant y(x) - P_2(x) \leqslant E.$$

令 $P_3(x) = \frac{1}{2}(P_1 + P_2)$ 则

$$-E \leqslant y(x) - P_3(x) \leqslant E,$$

因而 P_3 也是一个极小化极大多项式. 由题 22.21 必定有一个 $n+2$ 个点的序列, 在这些点上 $y - P_3(x)$ 交错地取 $\pm E$. 令 $P_3(x_+) = E$, 则在 x_+ 处我们有 $y - P_3 = E$, 或是

$$(y - P_1) + (y - P_2) = 2E.$$

因为左边没有一项能超过 E, 每项都必须等于 E, 因此 $P_1(x_+) = P_2(x_+)$. 类似地有 $P_1(x_-) = P_2(x_-)$. 因此多项式 P_1 及 P_2 在 $n+2$ 个点上偶合, 因而它们就是恒等的, 这就证明了极小化极大多项式对每个 n 的惟一性.

22.23 证明一个次数不大于 n 的多项式 $p(x)$, 对它而言误差 $y(x) - p(x)$ 在 $n+2$ 个点的一个集合上取交错的极值 $\pm e$, 它必定是极小化极大多项式.

证 ☞ 这将证明只有极小化极大多项式才能有这种等误差性质, 因而它对寻找和识别这类多项式是有用的. 我们有

$$\max |y(x) - p(x)| = e \geqslant E = \max |y(x) - P(x)|,$$

$P(x)$ 是惟一的极小化极大多项式. 假设 $e > E$. 此时由于

$$P - p = (y - p) + (P - y).$$

我们看到, 在 $y - p$ 的 $n+2$ 个极值点上, 量 $P - p$ 及 $y - p$ 有相同的符号. (右侧的第一项在这些点上等于 e 于是就超出第二项.) 但是 $y - p$ 的符号在这个集合上是交错的, 所以 $P - p$ 的符号也如此. 这全部有 $n+1$ 次变号, 它意味着 $P - p$ 有 $n+1$ 个零点. 由于 $P - p$ 是不大于 n 次的, 它必定恒等于零, 其结果为 $p = P$ 与 $E = e$. 这与我们的假设 $e > E$ 矛盾, 因而留给我们只有一种选择, 即 $e = E$. 该多项式 $p(x)$ 因而是(惟一的)极小化极大多项式 $P(x)$.

连续数据, 极小化极大多项式的例子

22.24 证明, 在区间 $(-1, 1)$ 上关于 $y(x) = x^{n+1}$ 的次数不大于 n 的极小化极大多项式可以通过将 x^{n+1} 表成一个 Chebyshev 多项式的和并且略去 $T_{n+1}(x)$ 项.

证 ☞ 令

$$x^{n+1} = a_0T_0(x) + \cdots + a_nT_n(x) + a_{n+1}T_{n+1}(x) = p(x) + a_{n+1}T_{n+1}(x),$$

这时误差为

$$E(x) = x^{n+1} - p(x) = a_{n+1}T_{n+1}(x).$$

我们还看到这个误差在 $T_{n+1}=\pm 1$ 的 $n+2$ 个点处有交错的极值 $\pm a_{n+1}$.这些点是 $x_k=\cos[k\pi/(n+1)]$, $k=0,1,\cdots,n+1$.比较上面二侧 x^{n+1} 的系数,我们还发现 $a_{n+1}=2^{-n}$.[$T_{n+1}(x)$的最高项系数为 2^n.参看题 21.42 及 21.43.]现在应用题 22.23 的结果并证明 $p(x)$为极小化极大多项式,以 $E=2^{-n}$.作为例证可以截断题 21.45 中的和来得到

$$n=1,\quad x^2\approx\frac{1}{2}T_0,\qquad 误差=\frac{T_2}{2};$$

$$n=2,\quad x^3\approx\frac{3}{4}T_1,\qquad 误差=\frac{T_3}{4};$$

$$n=3,\quad x^4\approx\frac{1}{8}(3T_0+4T_2),\qquad 误差=\frac{T_4}{8};$$

$$n=4,\quad x^5\approx\frac{1}{16}(10T_1+5T_3),\qquad 误差=\frac{T_5}{16}.$$

等等.注意在每种情况下极小化极大多项式(不大于 n 次的)实际上是 $n-1$ 次的.

22.25 证明在任何一个 Chebyshev 多项式级数 $\sum_{i=0}^{\infty} a_iT_i(x)$ 中,每一个部分和 S_n 是关于下一个和 S_{n+1}的次数不大于 n 的极小化极大多项式.[区间再次取$(-1,1)$.]

证 ☞ 正如与上题一样,只是取 $y(x)=S_{n+1}(x)$及 $p(x)=S_n(x)$,我们有

$$E(x) = S_{n+1}(x) - S_n(x) = a_{n+1}T_{n+1}(x).$$

再一次应用题 22.23 的结果.然而,还要注意到 $S_{n-1}(x)$可以不是次数不大于$n-1$的极小化极大多项式,因为 $a_nT_n+a_{n+1}T_{n+1}$不一定是一个等波纹函数.(不过在上题中依然是的,因为 a_n 为零.)

22.26 利用题 22.24 的结果,对于区间$(-1,1)$将多项式 $y(x)=x-\frac{1}{6}x^3+\frac{1}{120}x^5$ 缩减为一个三次多项式.

解 ☞ 它实际上是在题 21.50 中已完成,但是我们现在以一个新的眼光来看待这个结果.因为

$$x-\frac{1}{6}x^3+\frac{1}{120}x^5=\frac{169}{192}T_1-\frac{5}{128}T_3+\frac{1}{1920}T_5,$$

将 T_5 项截断留给我们的是关于 $y(x)$的不大于四次的极小化极大多项式,即

$$P(x)=\frac{169}{192}x-\frac{5}{128}(4x^3-3x),$$

它还只是近似地为关于 $\sin x$ 的同次的极小化极大多项式.进一步截断,把 T_3 项略去,不会产生关于 $y(x)$的极小化极大多项式,在任何情况下不是精确的.

22.27 在区间(a,b)上,对满足 $y''(x)>0$ 的函数 $y(x)$,寻找次数不大于 1 的极小化极大多项式.

解 ☞ 令该多项式为 $P(x)=Mx+B$.我们必定在(a,b)中找出 3 个点 $x_1<x_2<x_3$,在这些点上 $E(x)=y(x)-P(x)$以交错的符号达到它的极值.这就将 x_2 放到(a,b)的内部并要求 $E'(x_2)$为零或是 $y'(x_2)=M$.由于 $y''>0$,故 y'为严格增加的并且只能有一次与 M 相等,这意味着 x_2 能够是惟一的内部极值点.因此 $x_1=a$ 及 $x_3=b$.最后由等波纹性

$$y(a)-P(a)=-[y(x_2)-P(x_2)]=y(b)-P(b)$$

解之,我们得

$$M=\frac{y(b)-y(a)}{b-a},\qquad B=\frac{y(a)+y(x_2)}{2}-\frac{(a+x_2)[y(b)-y(a)]}{2(b-a)},$$

其中的 x_2 由 $y'(x_2)=[y(b)-y(a)]/b-a$ 所决定.

22.28 在区间$(0,\pi/2)$上应用上题于 $y(x)=-\sin x$.

解 ☞ 首先我们得到 $M=-2/\pi$;然后从 $y'(x_2)=M$,有 $x_2=\arccos(2/\pi)$.最后,

$$B = -\frac{1}{2}\sqrt{1-\frac{4}{\pi^2}} + \frac{1}{\pi}\arccos\frac{2}{\pi}.$$

并且从 $P(x) = Mx + B$ 我们得到

$$\sin x \approx \frac{2x}{\pi} + \frac{1}{2}\sqrt{1-\frac{4}{\pi^2}} + \frac{1}{\pi}\arccos\frac{2}{\pi},$$

这个逼近式就是极小化极大直线.

22.29　证明 $P(x) = x^2 + \frac{1}{8}$ 是在区间 $(-1,1)$ 上对 $y(x) = |x|$ 的极小化极大三次(或更低次的)逼近.

证　☞ 误差为 $E(x) = |x| - x^2 - \frac{1}{8}$ 并在 $x = -1, -\frac{1}{2}, 0, \frac{1}{2}, 1$ 上取极值 $-\frac{1}{8}, \frac{1}{8}, -\frac{1}{8}, \frac{1}{8}, \frac{1}{8}$. 这些在 $n+2=5$ 个点上具有符号交错并取大值为 $E = \frac{1}{8}$ 的误差保证了(由题 22.23) $P(x)$ 是一个次数 $n=3$ 或更低次的极小化极大多项式.

22.30　利用在区间 $(-1,1)$ 上的函数 $y(x) = e^x$ 作为例子来说明用来寻找极小化极大直线的交换法.

解　☞ 题 22.27 的方法会产生极小化极大直线,但是作为第一个简单的例证,我们暂且忽略那个方法并仿效题 22.5 中的过程来进行交换. 由于我们是在探求一条直线,所以我们需要具有最大误差为 $\pm E$, $n+2=3$ 个点. 尝试将 $x=-1,0,1$ 作为初始三元组,相应的 $y(x)$ 值约为 0.368, 1, 及 2.718. 易得关于这个三元组的等误差直线为

$$p_1(x) \approx 1.175x + 1.272,$$

在这三元组上具有误差 $h = \pm 0.272$. 在三元组之外在长度为 0.1 的诸区间上计算误差发现在 $x = 0.2$ 处有大小为 $H = 0.286$ 的最大误差(而且是负的). 据此我们形成一个新的三元组,把老的自变量 $x=0$ 换为新的 $x=0.2$. 它保留误差符号的交错,这是在早些时提供的交换法的第 4 步所要求的,现在我们就来仿效它. 在新的三元组上 $y(x)$ 近似地取 0.368, 1.221 及 2.718. 得到的等误差线为

$$p_2(x) = 1.175x + 1.264,$$

在这三元组上具有误差 $h = \pm 0.278$. 在三元组之外,预期最大误差在 $x=0.2$ 附近,我们以 0.01 的区间长检查这个邻域并且得到在 $x=0.16$ 处的误差为 0.279. 由于我们只用三位数进行计算,所以这是我们可以期望的最好的情况. 平移到三元组 $x=-1, 0.16, 1$ 后实际上将再次得到 $p_2(x)$.

现在让我们来考察题 22.27 的方法是怎样运作的. 以 $a=-1$ 及 $b=1$ 立刻得出 $M = (2.718 - 0.368)/2 = 1.175$. 这时方程 $y'(x_2) = e^{x_2} = 1.175$ 导出 $x_2 \approx 0.16$, 随后直接可得 $B = 1.264$ 的结果. 直线表示在下面的纵向尺度经过压缩的图 22.7 上.

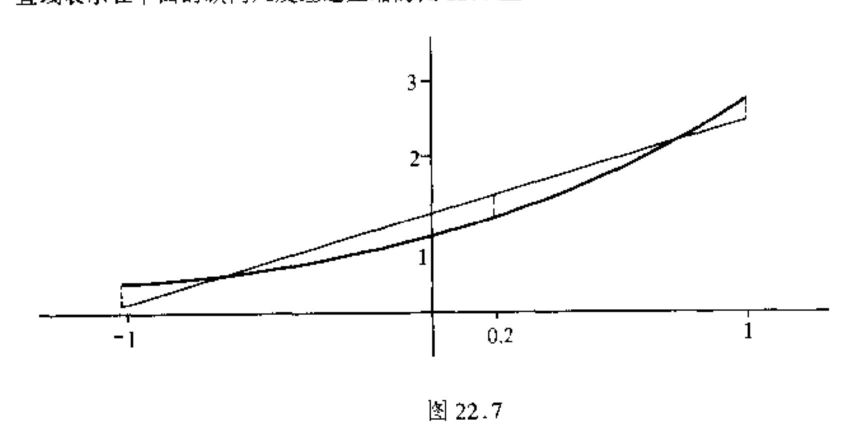


图 22.7

22.31　使用交换法寻找在 $(-1,1)$ 上 $y(x) = e^x$ 的极小化极大二次多项式.

解　☞ 回忆起 Chebyshev 多项式级数的截断通常导致相当于略掉第一项的近乎等波纹误差,故我们取 $T_3(x)$ 的 4 个极值点作为我们的初始四元组,它们是 $x = \pm 1, \pm\frac{1}{2}$. 交错地以 $\pm h$ 偏离下面 4 个点的抛物线,

x	-1	$-1/2$	$1/2$	1
e^x	0.3679	0.6065	1.6487	2.7183

证明在 $x=0.56$ 处有它的最大误差.新的四元组 $(-1,-0.5,0.56,1)$ 导出第二个抛物线在 $x=-0.44$ 处具有最大误差.下一个四元组为 $(-1,-0.44,0.56,1)$ 并证明为我们最终的那个.它的精确到5位小数的等波纹抛物线是

$$p(x)=0.55404x^2+1.13018x+0.98904,$$

而它的在四元组内外的最大误差则为 $H=0.04502$.

补 充 题

离散数据

22.32 证明对题22.2中的三个数据点的最小二乘直线为 $y(x)=\frac{1}{2}x-\frac{1}{6}$.证明它在数据点上的误差为 $\frac{1}{6},-\frac{1}{3},\frac{1}{6}$.所得到的Chebyshev直线为 $y(x)=\frac{1}{2}x-\frac{1}{4}$,带有误差 $-\frac{1}{4},\frac{1}{4},-\frac{1}{4}$.验证Chebyshev直线确有较小的最大误差而最小二乘直线具有较小的误差平方和.

22.33 对题21.2中的平均高尔夫记分应用交换法,产生极小化极大直线.用这条直线计算被平滑的平均记分.这个结果与那些由最小二乘方所得到的相比情况如何?

22.34 应用交换法于题21.5的数据,获得极小化极大直线以及相应的指数函数 $P(x)=Ae^{Mx}$.

22.35 推出对任意三元组 $(x_1,y_1),(x_2,y_2),(x_3,y_3)$ 的Chebyshev直线的公式 $y(x)=Mx+B$.这样的公式对编制交换法的机器计算程序是会有用的.

22.36 证明假如自变量 x_i 不是互异的,则极小化极大直线可能不是惟一确定的.例如,考虑三个点 $(0,0),(0,1)$ 及 $(1,0)$ 并证明所有介于 $y=\frac{1}{2}$ 及 $y=\frac{1}{2}-x$ 之间的直线都有 $H=\frac{1}{2}$.(参看图22.8.)

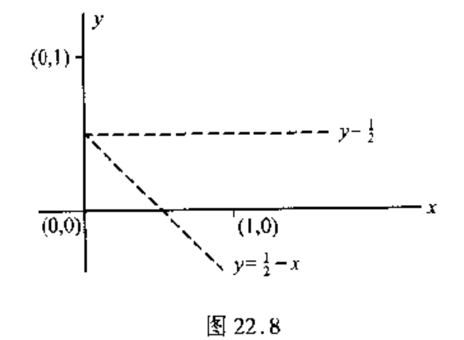


图22.8

22.37 对曲线 $y=\sin x$ 的4个点 $(0,0),\left(\pi/6,\frac{1}{2}\right),(\pi/3,\sqrt{3}/2)$ 及 $(\pi/2,1)$ 找出等误差抛物线.

22.38 对 $y=x^3,x=0\left(\frac{1}{4}\right)1$ 的5个点找出极小化极大抛物线.

22.39 对 $y=\cos x,x=0(\pi/12)\pi/2$ 的7个点使用交换法求得极小化极大抛物线.这抛物线的最大误差 $|h|$ 是什么?将它的精度与Taylor抛物线 $1-\frac{1}{2}x^2$ 的精度作比较.

22.40 推广交换法对 $y=\sin x,x=0(\pi/12)\pi/2$ 的7个点求得极小化极大三次多项式.这个三次式的最大误差 $|h|$ 是什么?将它的精度与Taylor三次式 $x-\frac{1}{6}x^3$ 的精度作比较.

连续数据

22.41 在区间 $(-1,1)$ 上对 $y(x)=x^6$ 找出次数不大于5的极小化极大多项式.误差是什么?

22.42 什么是关于 $y(x)=T_0+T_1+T_2+T_3$ 的次数不大于2的极小化极大多项式?它的误差是什么?通过证明其逼近误差不是等波纹的而证明 T_0+T_1 无论如何不是关于 $y(x)$ 的极小化极大直线.

22.43 对 $y(x)=1-\frac{1}{2}x^2+\frac{1}{24}x^4-\frac{1}{720}x^6$ 找出次数不大于5的极小化极大多项式,它的误差是什么?[区是是 $(-1,1)$.]

22.44 应用题22.27在区间 $(0,\pi/2)$ 上对 $y(x)=\cos x$ 求得极小化极大直线.

22.45 题22.27中的方法能否适用于在 $(-1,1)$ 上的 $y(x)=|x|$,或者是否由于 $y'(x)$ 中的间断使得方法

不能被应用?

22.46 在区间$(0,\pi/2)$上使用交换法来获得关于 $y=\cos x$ 的极小化极大直线,以三位小数进行计算并与题 22.44 中以其他方法所得结果进行比较.

22.47 在区间$(0,\pi/2)$上使用交换法来获得关于 $y=\cos x$ 的极小化极大抛物线.[你可能想用 $T_3(x)$的极值点,可通过一个变量变换,将它们转换到区间$(0,\pi/2)$上,作为一个初始四元组.]

22.48 找出一个最小次数的多项式它在区间$(0,\pi/2)$上以最大误差 0.005 逼近 $y(x)=\cos x$. 当然,舍入误差会限制精度,多项式可以视这个精度而定.

22.49 证明在所有的首项系数为 1 的 n 次多项式中,逼近 $f(x)=0$ 的极小化极大多项式是 $2^{1-n}T_n(x)$. 逼近区间取作$(-1,1)$. 这已经为题 22.17 到 22.23 所涵盖,但还是未施行下列历史上论证的细节. 令

$$p(x)=x^n+a_1x^{n-1}+\cdots+a_n$$

是任何一个具有所述类型的多项式. 由于 $T_n(x)=\cos(n\arccos x)$,我们有

$$\max|2^{1-n}T_n(x)|=2^{1-n}.$$

注意这个多项式在点 $x_k=\cos k\pi/n$, $k=0,1,\cdots,n$ 处交错地取它的极值 $\pm2^{1-n}$. 假设某多项式满足

$$\max|p(x)|<2^{1-n},$$

并令

$$P(x)=p(x)-2^{1-n}T_n(x),$$

这时 $P(x)$不大于 $n-1$ 次而且它不会恒等于零,因为这会要求 $\max|p(x)|=2^{1-n}$. 考虑值 $P(x_k)$. 由于 $p(x)$在这些点上是被 $2^{1-n}T_n(x)$所超过,故我们看到 $P(x_k)$有交错的符号. 由于连续性 $P(x)$ 因而必须有 n 个零点处在相继的x_k之间. 然而,对一个不恒等于零的次数不大于 $n-1$ 多项式而言这是不可能的. 这就证明了 $\max|p(x)|\geqslant 2^{1-n}$.

22.50 $y(x)=e^{(x+2)/4}$的值在下面的表中给出,寻找关于这个数据的极小化极大抛物线. 极小化极大误差是什么?

x	-2	-1	0	1	2
$y(x)$	1.0000	1.2840	1.6487	2.1170	2.7183

22.51 在区间$(-1,1)$上以不大于 0.005 的最大误差逼近 e^x 的多项式的最小次数是多少?

22.52 在区间$(0,1)$上 $\ln(1+x)$的 Taylor 级数收敛得如此地慢,为了达到 5 位精确需要几百项. 在同一区间上

$$p(x)=0.999902x-0.497875x^2+0.317650x^3$$
$$-0.193761x^4+0.085569x^5-0.018339x^6$$

的极大误差是什么?

22.53 以一个具有最小次数的多项式来逼近 $y(x)=1-x+x^2-x^3+x^4-x^5+x^6$, 在$(0,1)$中其误差不超过 0.005.

22.54 继续上题,产生一个最小次数的逼近多项式具有不超过 0.1 的误差.

第二十三章　有理函数逼近

配置

有理函数是多项式的商，因而构成比多项式要丰富得多的函数类.这个更大的函数类为精确逼近增加了希望.例如，带有极点的函数，简直不能期望在多项式方面的努力有好的回报，因为多项式并没有奇点.这种函数是有理逼近的主要目标.但是即使对非奇异函数，也存在宁愿选取有理逼近的可能性.

将要讨论两种类型的逼近，其过程类似于用于多项式逼近的那些.在预先给定点上的配置是选择有理逼近的一个基础，正像对于多项式那样.连分式和倒差分是所用的主要工具.所涉及的连分式取如下形式

$$y(x) = y_1 + \cfrac{x - x_1}{\rho_1 + \cfrac{x - x_2}{\rho_2 - y_1 + \cfrac{x - x_3}{\rho_3 - \rho_1 + \cfrac{x - x_4}{\rho_4 - \rho_2}}}},$$

如果需要的话它还可以继续下去.不大难发现这个特殊的分式可以重新整理为两个二次多项式的商，也就是一个有理函数.系数被称为倒差分，它们被选成要使得能实现配置.对眼前的这个例子，我们将得到

$$\rho_1 = \frac{x_2 - x_1}{y_2 - y_1}, \qquad \rho_2 - y_1 = \frac{x_3 - x_2}{\dfrac{x_3 - x_1}{y_3 - y_1} - \dfrac{x_2 - x_1}{y_2 - y_1}},$$

对 ρ_3 与 ρ_4 有类似的表达式.倒差分这个术语并没有不自然的地方.

极小化极大

极小化极大有理逼近在应用中同样赢得重要地位.它们的理论，包括等误差性以及一个交换算法，平行于多项式的情况，例如，能够找到一个有理函数

$$R(x) = \frac{1}{a + bx},$$

它交替地以 $\pm h$ 偏离三个指定的数据点 (x_i, y_i).当

$$\max | R(x_i) - y_i | = h$$

比用其他相同形式的有理函数来代替 $R(x)$ 所取的最大值为小时，$R(x)$ 是在这种意义下称作是对给定点的极小化极大有理函数.如果指定的点超过三个的话，则用交换算法来确定极小化极大 $R(x)$.它与极小化极大多项式的相似性是明显的.

Padé 逼近

它们具有形式

$$R_{mn}(x) = \frac{P_m(x)}{Q_n(x)}$$

其中 P_m 及 Q_n 分别是 m 及 n 次多项式.习惯上，规格化 $Q_n(0) = 1$.为逼近一个给定的函数 $y(x)$，Padé 提议使 y 与 R_{mn} 在某个指定点处它们的值连同头 N 次导数都相一致，其中 $N = m + n$.这就提供了 $N+1$ 个条件来决定 P_m 及 Q_n 中余下的 $N+1$ 个系数.问题中的点通常取作 $x=0$.倘若需要的话可以通过变量的一个适当平移来实现这一点.它与 $y(x)$ 在 $x=0$ 处的 Taylor 多项式的平行性是显然的，而事实上 Taylor 多项式是 R_{N0}.正如要看到的，对一个

给定的 N 通过选 $m=n+1$ 或 $m=n$. 可达到更高的精度，也就是说分子和分母的多项式的次数要选成大致相等.

题　解

配置有理函数

23.1　寻求有理函数 $y(x)=1/(a+bx)$，已知 $y(1)=1$ 及 $y(3)=\frac{1}{2}$.

解 ☞ 代入后要求 $a+b=1$ 及 $a+3b=2$，它迫使 $a=b=\frac{1}{2}$. 所要求的函数是 $y(x)=2/(1+x)$. 这个简单的问题说明这样的事实，即由配置来寻找一个有理函数等价于解一组关于未知系数的线性方程.

23.2　还要寻找有理函数 $y_2(x)=Mx+B$ 及 $y_3(x)=c+d/x$，它们也满足 $y(1)=1$ 及 $y(3)=\frac{1}{2}$，

解 ☞ 通过观察可以找到线性函数 $y_2(x)=(5-x)/4$. 对另外一个我们需要去满足系数方程 $c+d=1, 3c+d=\frac{3}{2}$ 而这意味着 $c=\frac{1}{4}, d=\frac{1}{4}$，使 $y_3(x)=(x+3)/4x$. 现在我们有三个有理函数，它们通过两个*给定的点. 肯定还有其他的. 然而在某种意义下这些是最简单的. 在 $x=2$ 处这三个函数提供给我们的插值为 $\frac{2}{3}$，$\frac{3}{4}$ 和 $\frac{5}{8}$. 在区间(1,3)内部三者在一定程度上彼此相近似. 在区间外它们有巨大的差异.（参看图 23.1）有理函数间的差异性超过多项式的，这非常有助于掌握所需要的有理函数类型的知识.

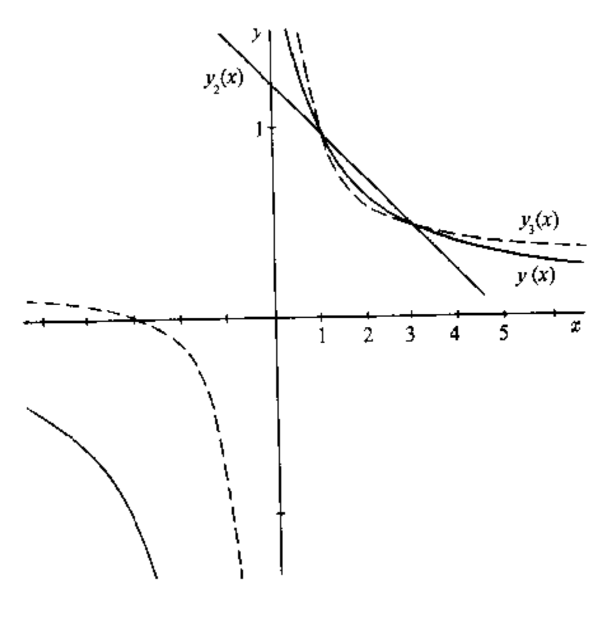


图 23.1

23.3　假设已知 $y(x)$ 的形式为 $y(x)=(a+bx^2)/(c+dx^2)$. 以要求 $y(0)=1, y(1)=\frac{2}{3}$，$y(2)=\frac{5}{9}$ 来决定 $y(x)$.

解 ☞ 代入后得到线性方程组

$$a=c,\qquad a+b=\frac{2}{3}(c+d),\qquad a+4b=\frac{5}{9}(c+4d).$$

---

* 译注：原文为三个.

由于包含的只是二个多项式的比，有一个系数可以取作 1，除非后来被证明它为零．尝试取 $d=1$．这时我们发现 $a=b=c=\frac{1}{2}$，因而 $y(x)=(1+x^2)/(1+2x^2)$．注意，有理函数 $y_2(x)=10/(10+6x-x^2)$ 也包含这三个点，$y_3(x)=(x+3)/[3(x+1)]$ 也如此．

连分式与倒差分

23.4　计算下面连分式在 $x=0,1$ 和 2 处的值：

$$y = 1 + \cfrac{x}{-3 + \cfrac{x-1}{-\frac{2}{3}}}.$$

解　☞ 直接计算表明 $y(0)=1$，$y(1)=\frac{2}{3}$ 和 $y(2)=\frac{5}{9}$．这些还是上题的值．这里的要点是这一类连分式之结构使这些值等于分式的逐次"渐近值"，即在 x 项之前和 $x-1$ 项之前截断分式所得到的部分值，当然还有末端的值．人们容易发现这个分式也可重新整理成我们的 $y_3(x)$．

23.5　在

$$y(x) = \frac{a_0 + a_1x + a_2x^2}{b_0 + b_1x + b_2x^2}$$

的情况下，推出有理函数与连分式之间的联系．

解　☞ 我们遵循另外一条历史途径．令 5 个数据点 (x_i, y_i)，$i=1,\cdots,5$ 为已知的．为了在这些点上的配置，对每对 x_i, y_i 有

$$a_0 - b_0y + a_1x - b_1xy + a_2x^2 - b_2x^2y = 0.$$

行列式方程

$$\begin{vmatrix} 1 & y & x & xy & x^2 & x^2y \\ 1 & y_1 & x_1 & x_1y_1 & x_1^2 & x_1^2y_1 \\ 1 & y_2 & x_2 & x_2y_2 & x_2^2 & x_2^2y_2 \\ 1 & y_3 & x_3 & x_3y_3 & x_3^2 & x_3^2y_3 \\ 1 & y_4 & x_4 & x_4y_4 & x_4^2 & x_4^2y_4 \\ 1 & y_5 & x_5 & x_5y_5 & x_5^2 & x_5^2y_5 \end{vmatrix} = 0$$

明显地具有所要求的性质．第二行通过下面的运算约化为 1,0,0,0,0,0：

从第二列减去第一列乘以 y_1．

从第四列减去第三列乘以 y_1．

从第六列减去第五列乘以 y_1．

从第五列减去第三列乘以 x_1．

从第三列减去第一列乘以 x_1．

此时行列式为下面的替代物所置换：

$$\begin{vmatrix} 1 & y-y_1 & x-x_1 & x(y-y_1) & x(x-x_1) & x^2(y-y_1) \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & y_2-y_1 & x_2-x_1 & x_2(y_2-y_1) & x_2(x_2-x_1) & x_2^2(y_2-y_1) \\ 1 & y_3-y_1 & x_3-x_1 & x_3(y_3-y_1) & x_3(x_3-x_1) & x_3^2(y_3-y_1) \\ 1 & y_4-y_1 & x_4-x_1 & x_4(y_4-y_1) & x_4(x_4-x_1) & x_4^2(y_4-y_1) \\ 1 & y_5-y_1 & x_5-x_1 & x_5(y_5-y_1) & x_5(x_5-x_1) & x_5^2(y_5-y_1) \end{vmatrix}.$$

按行列式的第二行展开然后：

以 $y-y_1$ 除第一行．

以 y_i-y_1 除第 i 行，对 $i=2,3,4,5$．

引入符号 $\rho_1(xx_1)=\frac{x-x_1}{y-y_1}$，方程现在可以写成

$$\begin{vmatrix} 1 & \rho_1(xx_1) & x & x\rho_1(xx_1) & x^2 \\ 1 & \rho_1(x_2x_1) & x_2 & x_2\rho_1(x_2x_1) & x_2^2 \\ 1 & \rho_1(x_3x_1) & x_3 & x_3\rho_1(x_3x_1) & x_3^2 \\ 1 & \rho_1(x_4x_1) & x_4 & x_4\rho_1(x_4x_1) & x_4^2 \\ 1 & \rho_1(x_5x_1) & x_5 & x_5\rho_1(x_5x_1) & x_5^2 \end{vmatrix} = 0.$$

现在重复这个运算，使第二行为 1,0,0,0,0：

从第二列减去第一列乘以 $\rho_1(x_2x_1)$.

从第四列减去第三列乘以 $\rho_1(x_2x_1)$.

从第五列减去第三列乘以 x_2.

从第三列减去第一列乘以 x_2.

这时行列式具有形式：

$$\begin{vmatrix} 1 & \rho_1(xx_1)-\rho_1(x_2x_1) & x-x_2 & x[\rho_1(xx_1)-\rho_1(x_2x_1)] & x(x-x_2) \\ 1 & 0 & 0 & 0 & 0 \\ 1 & \rho_1(x_3x_1)-\rho_1(x_2x_1) & x_3-x_2 & x[\rho_1(x_3x_1)-\rho_1(x_2x_1)] & x_3(x_3-x_2) \\ 1 & \rho_1(x_4x_1)-\rho_1(x_2x_1) & x_4-x_2 & x[\rho_1(x_4x_1)-\rho_1(x_2x_1)] & x_4(x_4-x_2) \\ 1 & \rho_1(x_5x_1)-\rho_1(x_2x_1) & x_5-x_2 & x[\rho_1(x_5x_1)-\rho_1(x_2x_1)] & x_5(x_5-x_2) \end{vmatrix}.$$

按第二行展开，然后

以 $\rho_1(xx_1)-\rho_1(x_2x_1)$除第一行.

以 $\rho_1(x_{i+1}x_1)-\rho_1(x_2x_1)$除第 i 行，对 $i=2,3,4$.

在这里是传统性再加的一步，目的为了保证要定义的量 ρ 的对称性质.(参看题 23.6)

以 y_1 乘第一列并加到第二列上.

以 y_1 乘第三列并加到第四列上.

引入符号 $\rho_2(xx_1x_2)=\dfrac{x-x_2}{\rho_1(xx_1)-\rho_1(x_2x_1)}+y_1$，现在方程简化成

$$\begin{vmatrix} 1 & \rho_2(xx_1x_2) & x & x\rho_2(xx_1x_2) \\ 1 & \rho_2(x_3x_1x_2) & x_3 & x_3\rho_2(x_3x_1x_2) \\ 1 & \rho_2(x_4x_1x_2) & x_4 & x_4\rho_2(x_4x_1x_2) \\ 1 & \rho_2(x_5x_1x_2) & x_5 & x_5\rho_2(x_5x_1x_2) \end{vmatrix} = 0.$$

另外的相似的简化产生

$$\begin{vmatrix} 1 & \rho_3(xx_1x_2x_3) & x \\ 1 & \rho_3(x_4x_1x_2x_3) & x_4 \\ 1 & \rho_3(x_5x_1x_2x_3) & x_5 \end{vmatrix} = 0,$$

其中

$$\rho_3(xx_1x_2x_3) = \frac{x-x_3}{\rho_2(xx_1x_2)-\rho_2(x_3x_1x_2)} + \rho_1(x_1x_2).$$

最终，最后的简化式为

$$\begin{vmatrix} 1 & \rho_4(xx_1x_2x_3x_4) \\ 1 & \rho_4(x_5x_1x_2x_3x_4) \end{vmatrix} = 0,$$

其中

$$\rho_4(xx_1x_2x_3x_4) = \frac{x-x_4}{\rho_3(xx_1x_2x_3)-\rho_3(x_4x_2x_3x_1)} + \rho_2(x_1x_2x_3).$$

我们推得 $\rho_4(xx_1x_2x_3x_4)=\rho_4(x_5x_1x_2x_3x_4)$. 刚刚引进的各个 ρ_i 称作 i 阶倒差分，而这四阶倒差分的等式等价于我们所开始的行列式方程，并且就是我们正在找的有理函数.

倒差分的定义现在以一个自然方式导出一个连分式.我们逐次地获得

$$y = y_1 + \frac{x-x_1}{\rho_1(xx_1)} = y_1 + \cfrac{x-x_1}{\rho_1(x_2x_1) + \cfrac{x-x_2}{\rho_2(xx_1x_2) - y_1}}$$

$$=y_1+\cfrac{x-x_1}{\rho_1(x_2x_1)+\cfrac{x-x_2}{\rho_2(x_3x_1x_2)-y_1+\cfrac{x-x_3}{\rho_3(xx_1x_2x_3)-\rho_1(x_1x_2)}}}$$

$$=y_1+\cfrac{x-x_1}{\rho_1(x_2x_1)+\cfrac{x-x_2}{\rho_2(x_3x_1x_2)-y_1+\cfrac{x-x_3}{\rho_3(x_4x_1x_2x_3)-\rho_1(x_1x_2)+\cfrac{x-x_4}{\rho_4(x_5x_1x_2x_3x_4)-\rho_2(x_1x_2x_3)}}}}$$

其中,在最后一个分母中,作为我们大量行列式简化之终点的四阶差分等式,最终被用上了.这就是那个使得上面的连分式成为所要求的有理函数的等式.(在所有这些计算背后的是这样的假设,即数据点确实属于这样一个有理函数,以后代数过程不会在某个点上中止.参看关于例外情况的例子.)

23.6　证明倒差分是对称的.

证　☞ 对一阶差分来说这是立刻就明白的,$\rho_1(x_1x_2)=\rho_1(x_2x_1)$.对二阶差分人们首先验证

$$\frac{x_3-x_2}{\dfrac{x_3-x_1}{y_3-y_1}-\dfrac{x_2-x_1}{y_2-y_1}}+y_1=\frac{x_3-x_1}{\dfrac{x_3-x_2}{y_3-y_2}-\dfrac{x_1-x_2}{y_1-y_2}}+y_2$$

$$=\frac{x_2-x_1}{\dfrac{x_2-x_3}{y_2-y_3}-\dfrac{x_1-x_3}{y_1-y_3}}+y_3,$$

由它可得在 $\rho_2(x_1x_2x_3)$中 x_i 可以用任何顺序进行排列.对更高阶的差分来说证明是类似的.

23.7　应用倒差分从表 23.1 的头两列中的数据 x,y 重新得到 $y(x)=1/(1+x^2)$.

表 23.1

x	y				
0	1				
		2			
①	$\frac{1}{2}$		-1		
		$-\frac{10}{3}$			
			$-\frac{1}{10}$	0	
2	$\frac{1}{5}$	-10			0
				40	
			$-\frac{1}{25}$		0
3	$\frac{1}{10}$				
		$-\frac{170}{7}$		140	
④	$\frac{1}{17}$		$-\frac{1}{46}$		
		$-\frac{442}{9}$			
5	$\frac{1}{26}$				

解　☞ 各阶倒差分也出现在这个表中.例如,40 这个项就是从带圈的项以下式得到的:

$$\rho_3(x_2x_3x_4x_5)=\frac{4-1}{\left(-\frac{1}{25}\right)-\left(-\frac{1}{10}\right)}+(-10)=40$$

$$=\frac{x_5-x_2}{\rho_2(x_3x_4x_5)-\rho_2(x_2x_3x_4)}+\rho_1(x_3x_4)$$

由题 23.5 中给出的定义,第三阶差分应该是

$$\rho_3(x_2x_3x_4x_5)=\frac{x_2-x_5}{\rho_2(x_2x_3x_4)-\rho_2(x_5x_3x_4)}+\rho_1(x_3x_4).$$

但是由对称性,这与我们上面用的是相同的.其他的差分可以用同样的方法获得.

连分式由顶对角线构成

$$y=1+\cfrac{x-0}{-2+\cfrac{x-1}{-1-1+\cfrac{x-2}{0-(-2)+\cfrac{x-3}{0-(-1)}}}}.$$

x	y				
1	30				
		$-\frac{1}{20}$			
2	10		$-\frac{10}{3}$		
		$-\frac{1}{5}$		$\frac{8}{5}$	
3	5		$-\frac{5}{3}$		0
		$-\frac{1}{2}$		1	
4	3		-3		
		0			
0	∞				

导出的连分式为

$$y = 30 + \cfrac{x-1}{-\cfrac{1}{20} + \cfrac{x-2}{-\cfrac{100}{3} + \cfrac{x-3}{\cfrac{33}{20} + \cfrac{x-4}{\cfrac{10}{3}}}}},$$

它约简成 $y(x)=60/[x(x+1)]$.

极小化极大有理函数

23.11 怎样才能找到一个有理函数 $R(x)=1/(a+bx)$，它偏离三个点 (x_1,y_1)，(x_2,y_2)，及 (x_3,y_3) 交错地为 $\pm h$？

解 ☞ 这三个条件为：

$$y_i - \frac{1}{a+bx_i} = h,\ -h,\ h \qquad 当\quad i=1,2,3.$$

可以写成

$$a(y_1-h)+b(y_1-h)x_1-1=0,$$
$$a(y_2+h)+b(y_2+h)x_2-1=0,$$
$$a(y_3-h)+b(y_3-h)x_3-1=0.$$

消去 a 及 b，我们得知 h 是由二次方程

$$\begin{vmatrix} y_1-h & (y_1-h)x_1 & -1 \\ y_2+h & (y_2+h)x_2 & -1 \\ y_3-h & (y_3-h)x_3 & -1 \end{vmatrix} = 0$$

所决定的. 选择有较小绝对值的根，我们将它回代而得到 a 及 b. (不难证明，实根总是存在的.)

23.12 应用题 23.11 的过程不这样三个点：$(0,0.83)$，$(1,1.06)$，$(2,1.25)$.

解 ☞ 二次方程变成 $4h^2-4.12h-0.130=0$，因而所要求的根为 $h=-0.03$. 于是系数 a 及 b 满足 $0.86a-1=0$，$1.03a+1.03b-1=0$ 从而它们为 $a\approx1.16$ 及 $b\approx-0.19$.

23.13 将上题加以推广，应用交换法来寻找一个极小化极大有理函数，其形式为 $R=1/(a+bx)$，对这些点：$(0,0.83)$，$(1,1.06)$，$(2,1.25)$，$(4,4.15)$.

解 ☞ 我们的问题是与先前的交换方法十分相似，上题的三元组当作一个初始三元组. 对这个三元组所获得的等误差有理函数为 $R_1(x)=1/(1.16-0.19x)$. 在这四个数据点上可以计算得它的误差为 -0.03，$+0.03$，-0.03，1.65. 我们发现 $R_1(x)$ 在 $x=4$ 处很差. 我们选择后三点为新的三元组，来保持交错的误差符号，新的二次方程为

$$6h^2-21.24h+1.47=0,$$

使 $h=0.07$. a，b 所满足的新方程为

$$a + b = 1.010, \qquad a + 2b = 0.758, \qquad a + 4b = 0.245,$$

使 $a \approx 1.265$ 及 $b \approx -0.255$. 在四个数据点上的误差现在为 0.04, 0.07, 0.07, 0.07; 而且由于我们现在的三元组中没有一个误差超过 0.07, 所以我们到此为止, 接受

$$R_2(x) = \frac{1}{1.265 - 0.255x}$$

为极小化极大逼近. 这是一个交换算法的典型展示. 我们的结果自然只精确到一位, 而数据本身只是两位精确, 所以进一步的努力看来是没有根据的. 有趣的是注意到该计算是十分敏感的. 例如, 在我们的 $R_2(x)$ 中舍去在第三位上那些 5, 就能改变 $R_2(4)$ 几乎半个单位. 这种敏感是出于极点在 $x = 5$ 附近. $R_1(x)$ 及 $R_2(x)$ 都在图 23.2 中给出.

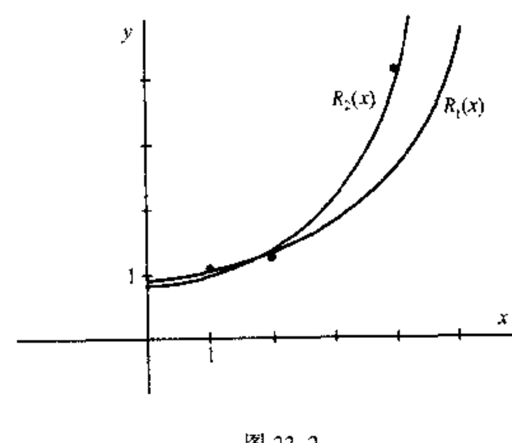


图 23.2

23.14 上题的数据点是对 $y(x) = 4/(5 - x)$ 加上一个不超过 5% 的随机"噪音"得来的. 用 $R_2(x)$ 来计算平滑值并与准确值以及原始数据进行比较.

解 ☞ 所要求的值如下, 还加上 $x = 3$ 处的值:

原始"噪音"数据	0.83	1.06	1.25	—	4.15
$R_2(x)$的值	0.79	0.99	1.32	2.00	4.08
$y(x)$的准确值	0.80	1.00	1.33	2.00	4.00

只有 $x = 4$ 处的误差是大的, 但它已经几乎被减少了一半. 在极点 $x = 5$ 处的影响是显然的. 以多项式进行的逼近会更加不成功.

23.15 对 Padé 有理函数

$$R_{mn}(x) = \frac{P_m(x)}{Q_n(x)},$$

其中

$$P_m(x) = a_0 + a_1x + a_2x^2 + \cdots + a_mx^m,$$

$$Q_n(x) = 1 + b_1x + b_2x^2 + \cdots + b_nx^n,$$

导出加在诸系数上的条件, 使得它当 $N = m + n$ 时满足

$$R_{mn}^{(k)}(0) = y^{(k)}(0) \qquad k = 0, 1, \cdots, N,$$

假设 $y(x)$ 有级数表达式

$$y(x) = c_0 + c_1x + c_2x^2 + \cdots.$$

解 ☞ 我们有

$$y(x) - R_{mn}(x) = \frac{\left(\sum_0^\infty c_ix^i\right)\left(\sum_0^n b_ix^i\right) - \sum_0^m a_ix^i}{\sum_0^n b_ix^i}.$$

如果右侧的分子没有比 x^{N+1} 低的项,我们就已经达到了所要的目标.为此我们需要

$$a_0 = b_0c_0, \qquad a_1 = b_0c_1 + b_1c_0, \qquad a_2 = b_0c_2 + b_1c_1 + b_2c_0.$$

一般地

$$a_j = \sum_{i=0}^{j} b_i c_{j-i}, \qquad j = 0,1,\cdots,N.$$

受 $b_0=1$ 约束的限制且

$$a_i = 0, \qquad 若\ i > m,$$

$$b_i = 0, \qquad 若\ i > n.$$

23.16 应用上题于 $y(x)=e^x$,取 $m=n=2$.

解 ☞ 对这个函数我们有 $c_0=1,c_1=1,c_2=\frac{1}{2},c_3=\frac{1}{6},c_4=\frac{1}{24}$导出的方程为

$$a_0 = 1, \qquad a_1 = 1 + b_1, \qquad a_2 = \frac{1}{2} + b_1 + b_2,$$

$$0 = \frac{1}{6} + \frac{1}{2}b_1 + b_2, \qquad 0 = \frac{1}{24} + \frac{1}{6}b_1 + \frac{1}{2}b_2.$$

它们的解是 $a_0=1,a_1=\frac{1}{2},a_2=\frac{1}{12},b_1=-\frac{1}{2}$以及 $b_2=\frac{1}{12}$,将它们回代,我们最终对 Padé 逼近有

$$R_{22}(x) = \frac{12+6x+x^2}{12-6x+x^2}.$$

在区间$(-1,1)$上它的绝对误差从中心处的零起到在 $x=1$ 处为 0.004.有意义的是注意到这种逼近反映了指数函数的一个基本性质,所以 $-x$ 替代x 便产生它的倒数.

23.17 对 $y(x)=e^x$ 显然

$$R_{40} = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots.$$

不过要用题 23.15 的方法来寻找 $R_{04}(x)$.

解 ☞ 适当的方程组包括 $a_0=1$,接下来三角形方程组

$$0 = 1 + b_1,$$

$$0 = \frac{1}{2} + b_1 + b_2,$$

$$0 = \frac{1}{6} + \frac{1}{2}b_1 + b_2 + b_3,$$

$$0 = \frac{1}{24} + \frac{1}{6}b_1 + \frac{1}{2}b_2 + b_3 + b_4.$$

导出逼近式

$$R_{04}(x) = \frac{1}{1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3 + \frac{1}{24}x^4},$$

这式子的分母是对 $y(x)$之倒数的一个 5 项逼近.这大概已经是所期望的.

相对于 R_{40}而言,在$(-1,1)$上 R_{04}在左半区间更接近 e^x,而在右半区间就离得更远了.R_{40}自始至终不如 R_{22}并且这是 Padé 逼近的普遍真理.m 等于或几乎等于n 的那些逼近是最精确的.

补　充　题

23.18 像在题 23.1 中那样,直接地求一个函数 $y(x)=1/(a+bx)$使 $y(1)=3$ 以及 $y(3)=1$.我们的连分式方法会产生这个函数吗?

23.19 直接求一个函数 $y(x)=1/(a+bx+cx^2)$满足 $y(0)=1,y(1)=\frac{1}{2}$以及 $y(10)=\frac{1}{4}$.我们的连分式方法能产生这个函数吗?

23.20 利用连分式方法来求一个有理函数具有下面的值:

x	0	1	2	3	4
y	-1	0	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{15}{17}$

23.21　利用连分式方法来求一个有理函数具有下面的值：

x	0	1	9	19
y	0	$\frac{1}{2}$	8.1	18.05

23.22　寻找一个有理函数具有这些值：

x	0	1	$+\infty$
y	$\frac{1}{2}$	$\frac{2}{3}$	1

23.23　寻找一个有理函数具有这些值：

x	0	1	2	4	∞
y	-2	$\pm\infty$	2	$\frac{6}{5}$	1

(记号$\pm\infty$涉及到一个极点在那儿函数改变符号.)

23.24　求一个具有下面给定值的有理函数.对$y(1.5)$进行插值.这个函数的"极点"在哪些地方?

x	0	± 1	± 2
y	$\frac{1}{2}$	1	$-\frac{1}{2}$

23.25　求关于函数$y(x)=x^2-1$在区间$(-1,1)$上的极小化极大函数

$$R(x)=\frac{1}{a+bx}.$$

23.26　在区间$(0,3)$上使用交换法求对$y(x)=e^x$的极小化极大逼近$R(x)=1/(a+bx)$.

23.27　对于点集(x_i,y_i),其中$i=1,\cdots,N$,推出一个交换法,用于寻求一个极小化极大逼近函数$R(x)=(a+bx)/(1+dx)$.将它应用于下面的数据：

x	0	1	2	3	4	5
y	0.38	0.30	0.16	0.20	0.12	0.10

使用$R(x)$来平滑y值.它们与$y(x)=\frac{1}{x+3}$接近的程度如何?这些带有随机误差的数据是源于这个函数的.

23.28　寻求一个有理函数它包含了这些点：

x	-1	0	1	2	3
y	∞	4	2	4	7

23.29　寻求一个有理函数它包含了这些点：

x	-2	-1	0	1	2
y	$-\infty$	0	3	8	∞

23.30　寻求一个有理函数包含了下面的点，这个函数有任何实极点吗？

x	−2	−1	0	1	2	3
y	$\frac{4}{3}$	2	2	$\frac{4}{3}$	$\frac{8}{7}$	$\frac{14}{13}$

23.31　用一个有理逼近函数以下面的表对 $y(1.5)$ 进行插值：

x	1	2	3	4
y	57.298677	28.653706	19.107321	14.335588

23.32　寻求一个有理函数，具有三次多项式除以二次多项式的形式，包含下面这些点：

x	0	1	2	3	4	5
y	12	0	−4	−6	6	4

23.33　以 $m=3, n=1$ 对题 23.16 进行工作.

23.34　以 $m=1, n=3$ 对题 23.16 进行工作.

第二十四章　三角逼近

离散数据

正弦函数和余弦函数它们具有多项式的许多好的性质、凭借着快速收敛的级数它们容易被计算.他们的逐次导数仍是正弦函数与余弦函数,对积分同样的情况也成立,它们也有正交的性质以及当然的周期性,这一点多项式并不具备.因而在逼近理论中这些熟悉的三角函数的使用就是可以理解的了.

一个配置预先给定的 $2L+1$ 个点三角和,它可以用

$$y(x)=\frac{1}{2}a_0+\sum_{k=1}^{L}\left(a_k\cos\frac{2\pi}{2L-1}kx+b_k\sin\frac{2\pi}{2L+1}kx\right)$$

这种形式来得到,若配置点为偶数个使用的形式稍有不同.这些正弦函数和余弦函数的一个正交性质是

$$\sum_{x=0}^{N}\sin\frac{2\pi}{N+1}jx\sin\frac{2\pi}{N+1}kx=\begin{cases}0, & 若\ j\neq k,\\ (N+1)/2, & 若\ j=k\neq 0.\end{cases}$$

$$\sum_{x=0}^{N}\sin\frac{2\pi}{N+1}jx\cos\frac{2\pi}{N+1}kx=0.$$

$$\sum_{x=0}^{N}\cos\frac{2\pi}{N+1}jx\cos\frac{2\pi}{N+1}kx=\begin{cases}0, & 若\ j\neq k,\\ (N+1)/2, & 若\ j=k\neq 0,N+1,\\ N+1, & 若\ j=k=0,N+1.\end{cases}$$

使其系数容易地被确定为

$$a_k=\frac{2}{2L+1}\sum_{x=0}^{2L}y(x)\cos\frac{2\pi}{2L+1}kx\qquad k=0,1,\cdots,L,$$

$$b_k=\frac{2}{2L+1}\sum_{x=0}^{2L}y(x)\sin\frac{2\pi}{2L+1}kx\qquad k=1,2,\cdots,L.$$

这些系数提供了具有指定形式的惟一的配置函数,对于偶数个配置点,譬如说 $2L$ 个相应的公式为

$$y(x)=\frac{1}{2}a_0+\sum_{k=1}^{L-1}\left(a_k\cos\frac{\pi}{L}kx+b_k\sin\frac{\pi}{L}kx\right)+\frac{1}{2}a_L\cos\pi x,$$

其中

$$a_k=\frac{1}{L}\sum_{x=0}^{2L-1}y(x)\cos\frac{\pi}{L}kx\qquad k=0,1,\cdots,L,$$

$$b_k=\frac{1}{L}\sum_{x=0}^{2L-1}y(x)\sin\frac{\pi}{L}kx\qquad k=1,\cdots,L-1.$$

对相同的离散数据进行最小二乘方逼近,使用相同类型的三角和.可简单地由将配置和截断而得到.这是一个既有名又方便的结果.正如在题 21.8 中观察到的那样,用正交函数的另外的表达式它也是成立的.在 $(2L+1)$ 个自变量的情况下,这里要被极小化的是

$$S=\sum_{x=0}^{2L}[y(x)-T_M(x)]^2,$$

其中 $T_M(x)$ 是缩短的和(M 比 L 小)

$$T_M(x)=\frac{1}{2}A_0+\sum_{k=1}^{M}\left(A_k\cos\frac{2\pi}{2L+1}kx+B_k\sin\frac{2\pi}{2L+1}kx\right).$$

刚才陈述的这个结果说明要将 S 极小化我们必须选取 $A_k=a_k$, $B_k=b_k$, S 的极小值可以表示为

$$S_{\min}=\frac{2L+1}{2}\sum_{k=M+1}^{L}(a_k^2+b_k^2)$$

当 $M=L$ 时它会变成零，这简直不是一种惊讶，因为从此我们再一次得到配置和.

周期性是三角和的一个明显的性质.如果一个数据函数不是基本上为周期的，它还是可以用于构造一个三角逼近，假如我们考虑的只是一个有限区间的话.给定的 $y(x)$ 于是可以想象为在区间外以一种方法将它延拓为周期的.

奇函数与偶函数通常用来作为一种延拓函数.一个奇函数具有性质 $y(-x)=-y(x)$.经典的例子是 $y(x)=\sin x$.对于一个具有周期 $P=2L$ 的奇函数，我们三角和的系数简化为

$$a_k = 0, \qquad b_k = \frac{4}{P}\sum_{x=1}^{L-1} y(x)\sin\frac{2\pi}{P}kx.$$

而一个偶函数则具有性质 $y(-x)=y(x)$.经典的例子为 $y(x)=\cos x$.对于一个周期为 $P=2L$ 的偶函数，系数变成

$$a_k = \frac{2}{P}[y(0)+y(L)\cos k\pi] + \frac{4}{P}\sum_{x=1}^{L-1} y(x)\cos\frac{2\pi}{P}kx, \qquad b_k = 0.$$

这些简化解释了奇函数和偶函数为什么会受欢迎.

连续数据

当供应的数据为连续的，Fourier 级数就替代了有限三角和，很多细节都是类似的，对于定义在 $(0,2\pi)$ 上的 $y(x)$，级数形式为

$$\frac{1}{2}\alpha_0 + \sum_{k=1}^{\infty}(\alpha_k\cos kt + \beta_k\sin kt).$$

正弦函数和余弦函数的第二种正交性质是

$$\int_0^{2\pi}\sin jt\,\sin kt\,dt = \begin{cases}0, & 若 j\neq k,\\ \pi, & j=k\neq 0.\end{cases}$$

$$\int_0^{2\pi}\sin jt\,\cos kt\,dt = 0.$$

$$\int_0^{2\pi}\cos jt\,\cos kt\,dt = \begin{cases}0, & 若 j\neq k,\\ \pi, & 若 j=k\neq 0,\\ 2\pi, & 若 j=k=0.\end{cases}$$

它的 Fourier 系数容易验明为

$$\alpha_k = \frac{1}{\pi}\int_0^{2\pi} y(t)\cos kt\,dt,$$

$$\beta_k = \frac{1}{\pi}\int_0^{2\pi} y(t)\sin kt\,dt.$$

由于级数具有 2π 的周期，我们必须将它的使用限止在给定区间 $(0,2\pi)$ 中，除非 $y(x)$ 碰巧有同样的周期.非周期函数在一个有限区间上也可以被接纳的，如果我们想象它们被延拓为周期的.奇和偶的延拓还是最常见的，在这种情况下，Fourier 系数大大地简化成像上面所述的.

Fourier 系数是与配置系数有关的.取一个奇数自变量的例子我们有，例如

$$a_j = \frac{1}{L}\left[\frac{1}{2}y(0) + \frac{1}{2}y(2L) + \sum_{x=1}^{2L-1} y(x)\cos\frac{\pi}{L}jx\right],$$

它就是对

$$\alpha_j = \frac{1}{L}\int_0^{2L} y(x)\cos\frac{\pi}{L}jx\,dx$$

的梯形法则逼近，其中用过一个变量变换为了得出这个类似的情形.

关于连续数据情况的最小二乘方逼近可以通过截断 Fourier 级数来得到，它将使积分

$$I = \int_0^{2\pi}[y(t) - T_M(t)]^2 dt.$$

极小化，其中

$$T_M = \frac{1}{2}A_0 + \sum_{k=1}^{M}(A_k\cos kt + B_k\sin kt).$$

换言之，要使 I 极小化我们必须选择 $A_k=\alpha_k, B_k=\beta_k$，$I$ 的极小值可以表示为

$$I_{\min} = \pi\sum_{k=M+1}^{\infty}(\alpha_k^2+\beta_k^2).$$

在对 $y(t)$ 作了十分弱的假设下出现平均收敛. 这意味着，当 M 趋于无穷时 $I_{\min}$ 的极限为零.

应用

在数值分析中三角函数逼近的两个主要应用是：

1. **数据平滑**. 由于通过截断可以这样方便地得到最小二乘逼近，这个应用看来是自然的，最小二乘原理平滑的效果类似于在多项式情况下观察到的那样.
2. **近似微分**. 同样在这里，隐约可见背景中的三角函数和的最小二乘外观有时应用一个诸如

$$y(x) \approx \frac{1}{10}[-2y(x-2) - y(x-1) + y(x+1) + 2y(x+2)]$$

的公式，它是由前面的最小二乘抛物线导出的，其结果通过三角函数和可得到进一步的平滑. 必须记住，这份平滑的危险性，它会抹掉目标函数的基本性质.

复数形式

前面所有的一切都可以表示为复数的形式，三角和变成

$$\sum_{j=-l}^{l}c_j e^{ijx}$$

其中 i 为虚数的单位，因为有 Euler 公式

$$e^{ix} = \cos x + i\sin x.$$

它等价于
$$\frac{a_0}{2} + \sum_{j=1}^{l}(a_j\cos jx + b_j\sin jx),$$

其中
$$a_j = c_j + c_{-j}, \qquad b_j = i(c_j - c_{-j}),$$

系数 $a_j, b_j c_j$ 可以是实数也可以是复数. Fourier 级数变成

$$f(x) = \sum_{j=-\infty}^{\infty} f_j e^{ijx}.$$

具有 Fourier 系数

$$f_j = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ijx}dx,$$

有限和

$$f_j^* = \frac{1}{N}\sum_{n=0}^{N-1} f(x_n)e^{-ijx_n},$$

其中 $x_n=2\pi n/N$ 对 $n=0$ 到 $N-1$，它是 f_j 的一个明显的逼近，也是在数据点 x_n 处插值 $f(x)$ 的三角函数和中适当的系数.

$$y(x) = \sum_{j=-l}^{l} f_j^* e^{ijx},$$

f_j^* 从根本上来说就是被称作离散 Fourier 变换的元素. 给定一个具有分量 v_0 到 v_{N-1} 向量 V，V 的离散 Fourier 变换可以定义为具有分量

$$v_j^T = \sum_{n=0}^{N-1} v_n\omega_N^{jn}$$

的向量 V^T，对 $j=0$ 到 $j=N-1$ 而 ω_N 为 1 的 N 次根.

$$\omega_N = e^{-2\pi i/N}$$

这些不同的关系式将在题解中得到考察.

它意味着通过使用离散变换来计算 Fourier 系数 f_j 的近似值是可能的,快速 Fourier 变换(FFT)使这种计算变得有效甚至当 N 值很大时,这些系数在许多应用中都是重要的.由于它们给出在一个复周期过程中分量项的相对分量.

题　解

以配置产生的三角函数和

24.1　证明正交条件当 $j+k\leqslant N$ 时

$$\sum_{x=0}^{N}\sin\frac{2\pi}{N+1}jx\ \sin\frac{2\pi}{N+1}kx=\begin{cases}0, & 若\ j\neq k\ 或\ j=k=0,\\ (N+1)/2, & 若\ j=k\neq 0.\end{cases}$$

$$\sum_{x=0}^{N}\sin\frac{2\pi}{N+1}jx\ \cos\frac{2\lambda}{N+1}kx=0.$$

$$\sum_{x=0}^{N}\cos\frac{2\pi}{N+1}jx\ \cos\frac{2\pi}{N+1}kx=\begin{cases}0, & 若\ j\neq k,\\ (N+1)/2, & 若\ j=k\neq 0,\\ N+1, & 若\ j=k=0.\end{cases}$$

证　☞ 它的证明可以通过初等三角.作为例证,

$$\sin\frac{2\pi}{N+1}jx\ \sin\frac{2\pi}{N+1}kx=\frac{1}{2}\left[\cos\frac{2\pi}{N+1}(j-k)x-\cos\frac{2\pi}{N+1}(j+k)x\right],$$

每个余弦函数和为零,由于所包含的角对称地分布在 0 与 2π 之间,而 $j=k\neq 0$ 除外,在这种情况下第一个余弦函数和为 $(N+1)/2$,另两个可用类似的模式加以证明.

24.2　关于奇数个自变量 $x=0,1,\cdots,N=2L$ 上的配置三角函数和可以取作

$$\frac{1}{2}a_0+\sum_{k=1}^{L}\left(a_k\cos\frac{2\pi}{2L+1}kx+b_k\sin\frac{2\pi}{2L+1}kx\right),$$

利用题 24.1 来决定系数 a_k 及 b_k.

解　☞ 为了得到 a_j 我们将 $y(x)$ 乘以 $\cos\frac{2\pi}{2L+1}jx$ 并求和,我们得到

$$a_j=\frac{2}{2L+1}\sum_{x=0}^{2L}y(x)\cos\frac{2\pi}{2L+1}jx,\qquad j=0,1,\cdots,L.$$

由于右侧所有其他项均为零.在 $y(x)$ 中的 1/2 因子使该结果对 $j=0$ 也成立.为了得到 b_j 我们将 $y(x)$ 乘以 $\sin\frac{2\pi}{2L+1}jx$ 并求和,得到

$$b_j=\frac{2}{2L+1}\sum_{x=0}^{2L}y(x)\sin\frac{2\pi}{2L+1}jx,\qquad j=1,2,\cdots,L.$$

因此只有一个这种表达式可以表示一个给定的函数 $y(x)$,系数由 $y(x)$ 在 $x=0,1,\cdots,2L$ 处的值惟一地确定.注意这个函数将以 $N+1$ 作为周期.

24.3　证明用题 24.2 的系数,当 $x=0,1,\cdots,2L$ 时三角和确等于 $y(x)$,这将证明在这些点上配置 $y(x)$ 的这种类型的和惟一存在.

证　☞ 暂时称这个和为 $T(x)$,并令 x^* 为 $2L+1$ 变量中的一个,将系数代入我们的公式导出

$$T(x^*)=\frac{2}{2L+1}\sum_{x=0}^{2L}y(x)\left[\frac{1}{2}+\sum_{k=1}^{L}\left(\cos\frac{2\pi}{2L+1}kx\ \cos\frac{2\pi}{2L+1}kx^*\right.\right.$$
$$\left.\left.+\sin\frac{2\pi}{2L+1}kx\ \sin\frac{2\pi}{2L+1}kx^*\right)\right]$$
$$=\frac{2}{2L+1}\sum_{x=0}^{2L}y(x)\left[\frac{1}{2}+\sum_{k=1}^{L}\cos\frac{2\pi}{2L+1}k(x-x^*)\right],$$

其中求和记号被交换过.现在最后一个和写成

$$\sum_{k=1}^{L}\cos\frac{2\pi}{2L+1}k(x-x^*)=\frac{1}{2}\sum_{k=1}^{L}\cos\frac{2\pi}{2L+1}k(x-x^*)$$

$$+\frac{1}{2}\sum_{k=L+1}^{2L}\cos\frac{2\pi}{2L+1}k(x-x^*).$$

因为余弦函数的对称性质这是可能的

$$\cos\frac{2\pi}{2L+1}k(x-x^*)=\cos\frac{2\pi}{2L+1}(2L+1-k)(x-x^*).$$

将 $k=0$ 项填入,现在我们得到

$$T(x^*)=\frac{1}{2L+1}\sum_{x=0}^{2L}y(x)\left[\sum_{k=0}^{2L}\cos\frac{2\pi}{2L+1}k(x-x^*)\right].$$

方括号中的项,当它变成 $2L+1$ 时由于正交性条件为零,除非 $x=x^*$. 因此 $T(x^*)=y(x^*)$,它就是要证明的.

24.4 假设已知 $y(x)$ 的周期为 3,寻找一个三角和,它包含了下面的数据点,并用它来插值 $y\left(\frac{1}{2}\right)$ 及 $y\left(\frac{3}{2}\right)$.

x	0	1	2
y	0	1	1

解 ☞ 使用题 24.2 的公式,我们得到

$$a_0=\frac{2}{3}(0+1+1)=\frac{4}{3},\qquad a_1=\frac{2}{3}\left(\cos\frac{2\pi}{3}+\cos\frac{4\pi}{3}\right)=-\frac{2}{3},$$

$$b_1=\frac{2}{3}\left(\sin\frac{2\pi}{3}+\sin\frac{4\pi}{3}\right)=0.$$

24.5 对自变量 x 为偶数个的 $(N+1=2L)$ 其配置和为

$$y(x)=\frac{1}{2}a_0+\sum_{k=1}^{L-1}\left(a_k\cos\frac{\pi}{L}kx+b_k\sin\frac{\pi}{L}kx\right)+\frac{1}{2}a_L\cos\pi x,$$

解 ☞ 具有配置点 $x=0,1,\cdots,N$. 通过一个与题 24.1 及题 24.2 几乎相同的论点得到系数为

$$a_j=\frac{1}{L}\sum_{x=0}^{2L-1}y(x)\cos\frac{\pi}{L}jx,\qquad j=0,1,\cdots,L,$$

$$b_j=\frac{1}{L}\sum_{x=0}^{2L-1}y(x)\sin\frac{\pi}{L}jx,\qquad j=1,\cdots,L-1.$$

函数 $y(x)$ 还是看作具有 $N+1$ 的周期. 将这些公式应用于下面的数据,然后计算 $y(x)$ 的极大值.

x	0	1	2	3
y	0	1	1	0

我们得到 $L=2$ 于是 $a_0=\frac{1}{2}(2)=1$, $a_1=\frac{1}{2}(-1)=-\frac{1}{2}$, $a_2=\frac{1}{2}(-1+1)=0$, $b_1=\frac{1}{2}(1)=\frac{1}{2}$. 因此三角函数和是

$$y(x)=\frac{1}{2}-\frac{1}{2}\cos\frac{1}{2}\pi x+\frac{1}{2}\sin\frac{1}{2}\pi x.$$

接着由标准过程得到 $y(x)$ 的极大值为

$$y\left(\frac{3}{2}\right)=\frac{1}{2}(1+\sqrt{2}).$$

用最小二乘方求三角函数和,离散数据

24.6 决定系数 A_k 与 B_k 使平方和

$$S=\sum_{x=0}^{2L}[y(x)-T_m(x)]^2=\text{极小},$$

其中 $T_m(x)$ 为三角函数和

$$T_m(x) = \frac{1}{2}A_0 + \sum_{k=1}^{M}\left(A_k\cos\frac{2\pi}{2L+1}kx + B_k\sin\frac{2\pi}{2L+1}kx\right),$$

而 $m < L$.

解 由于题 24.3 我们有

$$y(x) = \frac{1}{2}a_0 + \sum_{k=1}^{L}\left(a_k\cos\frac{2\pi}{2L+1}kx + b_k\sin\frac{2\pi}{2L+1}kx\right).$$

它们的差为

$$y(x) - T_m(x) = \frac{1}{2}(a_0 - A_0) + \sum_{k=1}^{M}\left[(a_k - A_k)\cos\frac{2\pi}{2L+1}kx\right.$$

$$\left. + (b_k - B_k)\sin\frac{2\pi}{2L+1}kx\right] + \sum_{k=M+1}^{L}\left[a_k\cos\frac{2\pi}{2L+1}kx + b_k\sin\frac{2\pi}{2L+1}kx\right].$$

将它平方,再对自变量 x 求和,并使用正交条件便得

$$S = \sum_{x=0}^{2L}[y(x) - T_m(x)]^2 = \frac{2L+1}{4}(a_0 - A_0)^2$$

$$+ \frac{2L+1}{2}\sum_{k=1}^{M}[(a_k - A_k)^2 + (b_k - B_k)^2] + \frac{2L+1}{2}\sum_{k=M+1}^{L}(a_k^2 + b_k^2).$$

只有头两项依赖于 A_k 及 B_k,并由于这些项均为非负的,极小值只可能以一种方法实现,即令这些项为零.因此对极小而言,有

$$A_k = a_k, \qquad B_k = b_k.$$

并且我们得到重要的结果,即将配置和 $T(x)$ 在 $k=M$ 处截断产生最小二乘方三角函数和 $T_M(x)$.(这实际上是在题 21.8 所得到的一般结果的另一种特殊情况.)我们还得到

$$S_{\min} = \frac{2L+1}{2}\sum_{k=M+1}^{L}(a_k^2 + b_k^2).$$

从一个几乎完全一样的计算证明

$$\sum_{x=0}^{2L}[y(x)]^2 = \sum_{x=0}^{2L}[T(x)]^2 = \frac{2L+1}{4}a_0^2 + \frac{2L+1}{2}\sum_{k=1}^{L}(a_k^2 + b_k^2).$$

它还可以表成

$$S_{\min} = \sum_{x=0}^{2L}[y(x)]^2 - \frac{2L+1}{4}a_0^2 - \frac{2L+1}{2}\sum_{k=1}^{M}(a_k^2 + b_k^2)$$

的形式.当 M 增加时这个和稳定地减少,当 $M=L$ 时达到零,从那时起最小二乘方与配置和恒等,对偶数个自变量 x 的情况一个多少有点类似的结果成立.

24.7 应用题 24.6 于题 24.4 的数据,取 $M=0$.

解 截断导致 $T_0(x) = \frac{2}{3}$.

奇的或偶的周期函数

24.8 假设 $y(x)$ 有周期 $P=2L$,即对所有的 x, $y(x+P)=y(x)$ 成立,证明此时题 24.5 中关于 a_j 和 b_j 的公式可以写成

$$a_j = \frac{2}{P}\sum_{x=-L+1}^{L}y(x)\cos\frac{2\pi}{P}jx, \qquad j = 0,1,\cdots,L,$$

$$b_j = \frac{2}{P}\sum_{x=-L+1}^{L}y(x)\sin\frac{2\pi}{P}jx, \qquad j = 1,\cdots,L-1.$$

解 由于正弦函数与余弦函数也有周期 P,使用变量 $x=0,\cdots,2L-1$ 或是变量 $-L+1,\cdots,L$ 并没有差别.任何这种 P 个连贯自变量的集合都会带来相同的系数.

24.9 假设 $y(x)$ 有周期 $P=2L$ 更加上还是一个奇函数,即 $y(-x) = -y(x)$.证明

$$a_j = 0, \qquad b_j = \frac{4}{P}\sum_{x=1}^{L-1}y(x)\sin\frac{2\pi}{P}jx.$$

证 ☞ 据周期性，$y(0)=y(P)=y(-P)$. 但是由于 $y(x)$是一个奇函数，$y(-P)=-y(P)$也成立. 这隐含了 $y(0)=0$. 以同样的方法我们得到 $y(L)=y(-L)=-y(L)=0$. 于是在关于 a_j 的和中，每一个在正 x 处的保留项与它的在负 x 处的对子相抵消，因此所有的 a_j 皆为零. 在有关 b_j 的和中，关于 x 的项与关于 $-x$ 的项是恒等的，所以我们在正的 x 上取和的二倍来得到 b_j.

24.10　对题 24.5 中的函数找出一个三角函数和，假设它延拓到一个周期为 $P=6$ 的奇函数.

解 ☞ 据前题所有 $a_j=0$，并且由于 $L=3$，得到

$$b_1=\frac{2}{3}\left(\sin\frac{\pi}{3}+\sin\frac{2\pi}{3}\right)=\frac{2}{\sqrt{3}},$$

$$b_2=\frac{2}{3}\left(\sin\frac{2\pi}{3}+\sin\frac{4\pi}{3}\right)=0.$$

从而使 $T(x)=(2/\sqrt{3})\sin(\pi x/3)$.

24.11　若 $y(x)$有周期 $P=2L$ 而且是个偶函数，即 $y(-x)=y(x)$，证明此时题 24.8 的公式变为

$$a_j=\frac{2}{P}[y(0)+y(L)\cos j\pi]+\frac{4}{P}\sum_{x=1}^{L-1}y(x)\cos\frac{2\pi}{P}jx,\quad j=0,1,\cdots,L,$$

$$b_j=0.$$

证 ☞ b_j 公式中关于 $\pm x$ 的项成对地抵消，在 a_j 公式中关于 $x=0$ 及 $x=L$ 的项可以像上面那样加以分开，在这之后剩下的项对 $\pm x$ 成为匹配的对子.

24.12　寻找关于题 24.5 的一个 $T(x)$假设它延拓到一个周期为 6 的偶函数(通过三角函数和它将造成该数据的三个表达式，只是形式不同. 参看题 24.5 及 24.10)

解 ☞ 所有的 b_j 都将为零，而取 $L=3$ 时我们得到 $a_0=\frac{4}{3}$, $a_1=0$, $a_2=-\frac{2}{3}$, $a_3=0$ 使得 $T(x)=\frac{2}{3}\left(1-\cos\frac{2}{3}\pi x\right)$.

连续数据，Fourier 级数

24.13　证明正交条件

$$\int_0^{2\pi}\sin jt\,\sin kt\,\mathrm{d}t=\begin{cases}0, & 若\ j\neq k,\\ \pi, & 若\ j=k\neq 0.\end{cases}$$

$$\int_0^{2\pi}\sin jt\,\cos kt\,\mathrm{d}t=0.$$

$$\int_0^{2\pi}\cos jt\,\cos kt\,\mathrm{d}t=\begin{cases}0, & 若\ j\neq k,\\ \pi, & 若\ j=k\neq 0,\\ 2\pi, & 若\ j=k=0.\end{cases}$$

其中 $j,k=0,1,\cdots$直至无穷.

证 ☞ 证明用的都是初等微积分. 例如

$$\sin jt\,\sin kt=\frac{1}{2}[\cos(j-k)t-\cos(j+k)t].$$

由于积分区间是余弦的一个周期，因而每个余弦积分为零. 当 $j=k\neq0$ 除外，在这种情况下第一个积分变成$\frac{1}{2}(2\pi)$. 另外二部分可以用类似的方式进行证明.

24.14　导出 Fourier 级数

$$y(t)=\frac{1}{2}\alpha_0+\sum_{k=1}^{\infty}(\alpha_k\cos kt+\beta_k\sin kt)$$

的系数公式

$$\alpha_j=\frac{1}{\pi}\int_0^{2\pi}y(t)\cos jt\,\mathrm{d}t,\qquad \beta_j=\frac{1}{\pi}\int_0^{2\pi}y(t)\sin jt\,\mathrm{d}t,$$

它们被称为 Fourier 系数. 事实上在正交函数的和或者级数中这类系数通常称为 Fourier 系数.

解 ☞ 证明按一条熟悉的途径, 以 $\cos jt$ 乘 $y(t)$ 并在整个区间 $(0,2\pi)$ 上积分, 右边所有各项除一项外均为零, 于是就出现关于 α_j 的公式. 在 α_0 项前的因子 $\frac{1}{2}$ 使该结果对 $j=0$ 也成立. 为了得到 β_j 我们以 $\sin jt$ 乘 $y(t)$ 并积分. 这里我们作了级数收敛于 $y(t)$ 以及逐次积分成立的假设, 它在关于 $y(t)$ 光滑性的十分弱的假设下, 按 Fourier 级数的理论被证明的, 明显地, $y(t)$ 肯定也有周期 2π.

24.15 求得关于 $y(t)=|t|,\ -\pi\leqslant t\leqslant\pi$ 的 Fourier 级数.

解 ☞ 令 $y(t)$ 被延拓为一个周期为 2π 的偶函数. (参看图 24.1 中的实线). 在我们的系数公式中积分限可以移至 $(-\pi,\pi)$, 我们发现所有 $\beta_j=0$. 同样 $\alpha_0=\pi$; 而对于 $j>0$

$$\alpha_j=\frac{2}{\pi}\int_0^\pi t\cos jt\,\mathrm{d}t=\frac{2(\cos j\pi-1)}{\pi j^2}.$$

因此

$$y(t)=\frac{\pi}{2}-\frac{4}{\pi}\left(\cos t+\frac{\cos 3t}{3^2}+\frac{\cos 5t}{5^2}+\cdots\right).$$

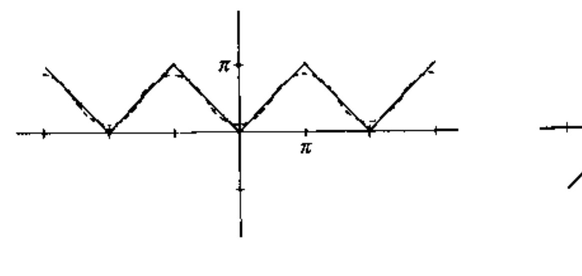


图 24.1

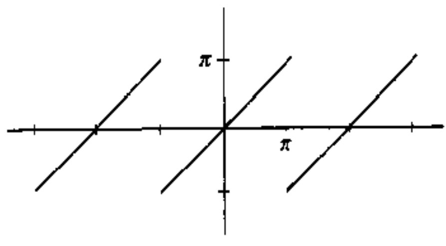


图 24.2

24.16 求得关于 $y(t)=t,\ -\pi<t<\pi$ 的 Forier 级数.

解 ☞ 我们将 $y(t)$ 延拓到一个具有周期 2π 的奇函数. (参看图 24.2). 再一次将积分限移至 $(-\pi,\pi)$, 我们得到所有 $\alpha_j=0$, 以及

$$\beta_j=\frac{2}{\pi}\int_0^\pi t\sin jt\,\mathrm{d}t=\frac{2(-1)^{j-1}}{j}.$$

因此

$$y(t)=2\left(\sin t-\frac{\sin 2t}{2}+\frac{\sin 3t}{3}-\frac{\sin 4t}{4}+\cdots\right).$$

注意题 24.15 的余弦级数比正弦级数收敛更快. 这与 $y(t)$ 在那题中是连续的而在本题中不是的有关. $y(t)$ 越光滑收敛就越快, 还注意到在不连续的点上我们的正弦级收敛于零, 它是 $y(t)$ 的左、右极值 (π 及 $-\pi$) 的平均.

24.17 寻找关于 $y(t)=\begin{cases}t(\pi-t), & 0\leqslant t\leqslant\pi,\\ t(\pi+t), & -\pi\leqslant t\leqslant 0.\end{cases}$ 的 Fourier 级数.

解 ☞ 将该函数延拓成一个周期为 2π 的奇函数, 我们将结果展示在图 24.3 中. 注意到这个函数没有角点, 在 $t=0$ 处它的左导数与右导数均为 π, 而 $y'(\pi)$ 及 $y'(-\pi)$ 则均为 $-\pi$, 因而即使是延拓的周期函数都没有角点. 这个额外的平滑化会反映在 Fourier 系数上. 使用积分限 $(-\pi,\pi)$ 我们再一次得到所有 $\alpha_j=0$, 以及

$$\begin{aligned}\beta_j&=\frac{2}{\pi}\int_0^\pi t(\pi-t)\sin jt\,\mathrm{d}t\\&=\frac{2}{\pi}\int_0^\pi\frac{\pi-2t}{j}\cos jt\,\mathrm{d}t\\&=\frac{4}{\pi j^2}\int_0^\pi\sin jt\,\mathrm{d}t=\frac{4(1-\cos j\pi)}{\pi j^3}.\end{aligned}$$

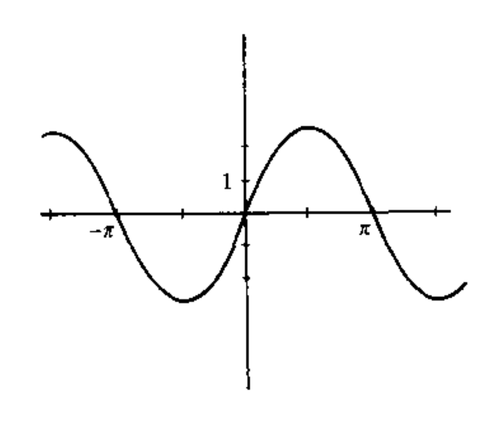


图 24.3

因而这个级数是

$$y(t)=\frac{8}{\pi}\left(\sin t+\frac{\sin 3t}{3^3}+\frac{\sin 5t}{5^3}+\cdots\right).$$

系数如负3次方似地减少,它造成了十分满意的收敛.函数的额外的平滑化证明为有用的.

24.18 证明对于 Bernoulli 函数

$$F_n(x)=B_n(x),\qquad 0<x<1,\qquad F_n(x\pm m)=F_n(x).\qquad m\text{ 为整数}$$

$B_n(x)$为一个 Bernoulli 多项式,当 n 为偶数时 Fourier 级数为

$$F_n(x)=(-1)^{(n/2)+1}n!\left[\frac{2}{(2\pi)^n}\right]\sum_{k=1}^{\infty}\frac{\cos 2\pi kx}{k^n},$$

当 n 为奇数时 Fourier 级数为

$$F_n(x)=(-1)^{(n+1)/2}n!\left[\frac{2}{(2\pi)^n}\right]\sum_{k=1}^{\infty}\frac{\sin 2\pi kx}{k^n},$$

这个结果曾在和与级数这一章的题 17.28 中用过.

证 ☞ 由于 $B_1(x)=x-\frac{1}{2}$,关于 $F_1(x)$的级数可以从系数公式直接求得,它为

$$F_1(x)=-\frac{1}{\pi}\left(\frac{\sin 2\pi x}{1}+\frac{\sin 4\pi x}{2}+\frac{\sin 6\pi x}{3}+\cdots\right).$$

通过积分,并回忆

$$B_n'(x)=nB_{n-1}(x),\qquad \int_0^1 B_n(x)dx=0,\qquad \text{当 } n>0,$$

我们立即得到

$$F_2(x)=\frac{2\cdot 2!}{(2\pi)^2}\left(\frac{\cos 2\pi x}{1}+\frac{\cos 4\pi x}{2^2}+\frac{\cos 6\pi x}{3^2}+\cdots\right).$$

接下来的积分使

$$F_3(x)=\frac{2\cdot 3!}{(2\pi)^3}\left(\frac{\sin 2\pi x}{1}+\frac{\sin 4\pi x}{2^3}+\frac{\sin 6\pi x}{3^3}+\cdots\right).$$

可以用一个归纳法来完成一个形式上的证明,在这里知道对一个 Fourier 级数逐项积分通常会产生被积分函数的 Fourier 级数是有好处的.类似的陈述对微分来说一般不成立.关于细节可参看 Fourier 级数的理论处理.

24.19 题 24.5,或是题 24.2 的配置系数与题 24.14 的 Fourier 系数有何关系?

解 ☞ 有许多方法可用来作这个比较,最有兴趣的是注意到在题 24.5 中的那个,假设 $y(x)$有周期 $P=2L$,我们可以将 a_j 改写为

$$a_j=\frac{1}{L}\left[\frac{1}{2}y(0)+\frac{1}{2}y(2L)+\sum_{x=1}^{2L-1}y(x)\cos\frac{\pi}{L}jx\right].$$

而这就是逼近于 Fourier 系数的梯形法则

$$a_j=\frac{1}{\pi}\int_0^{2\pi}y(t)\cos jt\,dt=\frac{1}{L}\int_0^{2L}y(x)\cos\frac{\pi}{L}jx\,dx$$

类似结果对 b_j,β_j 以及对题 24.2 中的系数也成立.由于当 L 变成无穷时梯形公式收敛于积分,我们发现配置系数收敛到 Fourier 系数(这里为了方便起见我们可以将周期固定为 2π).对 Chebyshev 多项式的有关类推可以参看题 21.53 到题 21.55.

最小二乘方,连续数据

24.20 决定系数 A_k 及 B_k 使积分

$$I=\int_0^{2\pi}[y(t)-T_M(t)]^2dt$$

为极小,其中 $T_M(t)=\frac{1}{2}A_0+\sum_{k=1}^{M}(A_k\cos kt+B_k\sin kt)$.

解 ☞ 多少有点像题 24.6 中那样,我们首先求得

$$y(t)-T_M(t)=\frac{1}{2}(\alpha_0-A_0)+\sum_{k=1}^{M}[(\alpha_k-A_k)\cos kt+(\beta_k-B_k)\sin kt]$$

$$+\sum_{k=M+1}^{\infty}(\alpha_k\cos kt+\beta_k\sin kt),$$

然后将它平方，再积分，并用正交条件来得到

$$I=\frac{\pi}{2}(\alpha_0-A)^2+\pi\sum_{k=1}^{M}[(\alpha_k-A_k)^2+(\beta_k-B_k)^2]+\pi\sum_{k=M+1}^{\infty}(\alpha_k^2+\beta_k^2).$$

对于一个极小值我们选所有 $A_k=\alpha_k, B_k=\beta_k$，因而

$$I_{\min}=\pi\sum_{k=M+1}^{\infty}(\alpha_k^2+\beta_k^2).$$

又一次我们得到重要结果，Fourier 级数在 $k=M$ 处的截断产生最小二乘方和 $T_M(t)$.（它又一次地是题 21.8 的一个特殊情况.）极小积分可以改写成

$$I_{\min}=\int_0^{2\pi}[y(t)]^2\mathrm{d}t-\frac{1}{2}\pi\alpha_0^2-\pi\sum_{k=1}^{M}(\alpha_k^2+\beta_k^2),$$

当 M 增加时，它减小，而它在 Fourier 级数的理论中被证明当 M 变成无穷时 $I_{\min}$ 趋于零. 这被称为平均收敛.

24.21 对题 24.15 中的函数寻找最小二乘方和，取 $M=1$.

解 ☞ 截断带来 $T_1(t)=\pi/2-(4/\pi)\cos t$. 这个函数在图 24-1 中以实线标出，注意它磨光了 $y(t)$ 的角.

借助 Fourier 分析的平滑化

24.22 Fourier 分析方法磨光数据的基础是什么？

解 ☞ 假如我们想象给定的数值数据是由函数的真值带有迭加的随机误差，真函数相对地光滑而迭加上去的误差十分不光滑，然后在题 24.15 到 24.17 的例子提示了一种将函数部分地从误差中分开的方法. 由于真函数是光滑的，它的 Fourier 系数将快速地减少. 但是误差的不光滑性提示了它的 Fourier 系数可能减少得十分慢，假如果真如此. 因此，超过某一个地方，这综合的级数几乎完全由误差组成. 假如我们简单地将级数在正确的地方加以截断，这样，我们就可甩掉大部分误差. 在保留的项中当然仍有误差的作用. 由于截断产生一个最小二乘方逼近，我们也可以把这个方法看成是最小二乘方平滑.

24.23 应用前题中的方法于下面的数据：

x	0	1	2	3	4	5	6	7	8	9	10
y	0	4.3	8.5	10.5	16.0	19.0	21.1	24.9	25.9	26.3	27.8

x	11	12	13	14	15	16	17	18	19	20
y	30.0	30.4	30.6	26.8	25.7	21.8	18.4	12.7	7.1	0

解 ☞ 假如函数在二个端点处为真正的零，我们可以假设它被延拓到一个周期为 $P=40$ 的奇函数. 这样的一个函数甚至有连续的一阶导数，它有助于加速 Fourier 级数的收敛，使用题 24.9 的公式，我们现在来计算 b_j.

j	1	2	3	4	5	6	7	8	9	10
b_j	30.04	−3.58	1.35	−0.13	−0.14	−0.43	0.46	0.24	−0.19	0.04

j	11	12	13	14	15	16	17	18	19	20
b_j	0.34	0.19	0.20	−0.12	−0.36	−0.18	−0.05	−0.37	0.27	

它快速地减少是明显的,我们可以把所有远于头三项或四项的 b_j 看作大部分是误差的作用.假如使用了四项,我们得到三角函数和

$$T(x) = 30.04\sin\frac{\pi x}{20} - 3.58\sin\frac{2\pi x}{20} + 1.35\sin\frac{3\pi x}{20} - 0.13\sin\frac{4\pi x}{20}.$$

该和的值可以与原始数据进行比较,它们是 $y(x)=x(400-x^2)/100$ 受到人为地引进的随机误差污染的实际所得的值(参看表 24.1).给定数值的 RMS 误差为 1.06,而平滑过的数据其 RMS 误差为 0.80.

表 24.1

x	给定的	准确的	磨光的	x	给定的	准确的	磨光的
1	4.3	4.0	4.1	11	30.0	30.7	29.5
2	8.5	7.9	8.1	12	30.4	30.7	29.8
3	10.5	11.7	11.9	13	30.6	30.0	29.3
4	16.0	15.6	15.5	14	26.8	28.6	28.0
5	19.0	18.7	18.6	15	25.7	26.2	25.8
6	21.1	22.7	21.4	16	21.8	23.0	22.4
7	24.9	24.6	23.8	17	18.4	18.9	18.0
8	25.9	26.9	25.8	18	12.7	13.7	12.6
9	26.3	28.7	27.4	19	7.1	7.4	6.5
10	27.8	30.0	28.7	20			

24.24 在同样给定的数据的基础上逼近前题中函数的导数 $y'(x)=(400-3x^2)/100$.

解 ☞ 首先我们将应用公式

$$y'(x) \approx \frac{1}{10}[-2y(x-2) - y(x-1) + y(x+1)^* + 2y(x+2)]$$

它是前面由关于 5 个自变量 $x-2,\cdots,x+2$ 的最小二乘抛物线导出的,对 4 个端点自变量可用类似的公式,其结果形成表 24.2 的第二列,使用这个局部最小二乘抛物线已经相当于对原始数据 x, y 的局部平滑,我们现在尝试进一步地以 Fourier 方法进行全局平滑,由于奇函数的导数是偶函数,题 24.11 的公式是适当的.

$$a_j = \frac{1}{20}[y'(0) + y'(20)\cos j\pi] + \frac{1}{10}\sum_{x=1}^{19} y'(x)\cos\frac{\pi}{20}jx,$$

这些系数可以计算为:

j	0	1	2	3	4	5	6	7	8	9	10
a_j	0	4.81	−1.05	0.71	−0.05	0.05	−0.20	0.33	0.15	0.00	0.06

j	11	12	13	14	15	16	17	18	19	20
a_j	0.06	0.06	−0.03	0.11	0.06	0.14	−0.04	0.16	−0.09	0.10

这个严重的下降仍是值得注意的.忽略掉所有超过 $j=4$ 的项,我们得到

* 译注:原书此处为 $(x-1)$.

$$y'(x) \approx 4.81\cos\frac{\pi x}{20} - 1.05\cos\frac{2\pi x}{20} + 0.71\cos\frac{3\pi x}{20} - 0.05\cos\frac{4\pi x}{20}.$$

对 $x=0,\cdots,20$ 计算这个公式便产生表 24.2 中的第三列,最后一列给出准确值.第二列中的 RMS 误差,在以最小二乘抛物线进行局部平滑后为 0.54,而在附加的 Fourier 平滑后所得第三列中的 RMS 误差为 0.39.

表 24.2

x	局部	Fourier	准确	x	局部	Fourier	准确
0	5.3	4.4	4.0	11	1.1	0.5	0.4
1	4.1	4.4	4.0	12	-0.1	-0.1	-0.3
2	3.8	4.1	3.9	13	-1.2	-0.9	-1.1
3	3.7	3.8	3.7	14	-2.2	-1.8	-1.9
4	3.4	3.4	3.5	15	-2.9	-2.9	-2.8
5	3.4	3.0	3.2	16	-3.6	-4.0	-3.7
6	2.6	2.5	2.9	17	-4.6	-5.0	-4.7
7	1.9	2.1	2.5	18	-5.5	-5.8	-5.7
8	1.5	1.8	2.1	19	-7.1	-6.4	-6.8
9	1.2	1.4	1.6	20	-6.4	-6.6	-8.0
10	1.3	1.0	1.0				

复数形式

24.25 当 j 与 k 为整数时,对函数 e^{ijx},e^{ikx} 证明以下的正交性质.上面的一横表示复共轭.

$$\int_0^{2\pi} \overline{e^{ijx}} e^{ikx} dx = \begin{cases} 0, & 若\ k \neq j, \\ 2\pi, & 若\ k = j. \end{cases}$$

证 ☞ 证明是初等的,该积分当 $k\neq j$ 时立刻还原为

$$\int_0^{2\pi} e^{i(k-j)} dx = \frac{1}{i(k-j)} e^{i(k-j)x} \Big|_0^{2\pi}.$$

但是它在上下限处均等于 1,因此为零,当 $k=j$,上面的左端明显地为 2π.

24.26 导出复数形式的 Fourier 系数公式.

解 ☞ 证明取一个熟悉的途径.Fourier 级数为

$$f(x) = \sum_{j=-\infty}^{\infty} f_j e^{ijx}$$

乘以 e^{-ikx} 并积分便得到

$$\int_0^{2\pi} f(x) e^{-ikx} dx = \sum_{j=-\infty}^{\infty} f_j e^{-ikx} e^{ijx} dx$$

而且由于右边所有的项除 $j=k$ 项外均因正交性而为零,便得所要的结果.

$$f_k = \frac{1}{2\pi}\int_0^{2\pi} f(x) e^{-ikx} dx$$

24.27 证明函数 e^{ijx_n},e^{ikx_n} 在下面意义下为正交,即

$$\sum_{n=0}^{N-1} \overline{e^{ijx_n}} e_n^{ikx} = \begin{cases} N, & 若\ k = j, \\ 0, & 若\ k \neq j. \end{cases}$$

如前,这里 $x_n = 2\pi n/N$.

证 ☞ 我们将得到一个以 $r = e^{i(k-j)2\pi/N}$ 为公比的几何和,

$$\sum_{n=0}^{N-1} \overline{e^{ijx_n}} e^{ikx_n} = \sum_{n=0}^{N-1} e^{i(k-j)x_n} = e^{i(k-j)x_0}(1 + r + r^2 + \cdots + r^{N-1}).$$

当 $j=k$ 时我们有 $r=1$ 而该和就是 N. 在其他的情况下,可以熟悉的公式求得该幂的和为 $(1-r^N)/(1-r)$. 但 r^N 就是 $e^{2\pi i(k-j)}$. 它等于 1, 使分子为零因而确立正交性.

24.28 证明若 $N=2l+1$ 则三角函数和

$$\sum_{j=-l}^{l} d_j e^{ijx}$$

必定有系数 $d_j = f_j^*$, 假如它是用来配置在点 $x_n = 2\pi n/N$ 处的 $f(x)$ 的话.

证 ☞ 假设配置存在,乘以 $\overline{e^{ikx_n}}$ 再加起来

$$\sum_{n=0}^{N-1} f(x_n)\overline{e^{ikx_n}} = \sum_{n=0}^{N-1} \overline{e^{ikx_n}} \sum_{j=-l}^{l} d_j e^{ijx_n} = \sum_{j=-l}^{l} d_j \sum_{n=0}^{N-1} \overline{e^{ikx_n}} e^{ijx_n}$$

还是这样,右侧的所有项除一项 $j=k$ 外为零,因而我们有

$$\sum_{n=0}^{N-1} f(x_n) e^{ikx_n} = d_k(N) = f_k^* N.$$

24.29 系数 f_j^* 与离散 Fourier 变换有什么关系?

解 ☞ 令 V 为具有分量 $f(x_0), \cdots, f(x_{N-1})$ 的向量,当 $N=2l+1$ 时,它使 V 为 $(2l+1)$ 维的,正像三角函数和 $\sum_{j=-l}^{l} f_j^* e^{ijx}$ 系数 f_j^* 向量的维数.

其中

$$f_j^* = \frac{1}{N}\sum_{n=0}^{N-1} f(x_n) e^{-ijx_n}.$$

当 $j=-l$ 到 $j=l$, 将它与

$$v_j^T = \sum_{n=0}^{N-1} v_n \omega_N^{nj} = \sum_{n=0}^{N-1} f(x_n) e^{-ijx_n}$$

相比,其中 $x_n = 2\pi n/N$ 且 $j=0$ 到 $j=N-1$, 吻合是显著的. 我们确有一问题:即成立的范围并不偶合,但是我们可以推导在何处范围相重,从 $j=0$ 到 $j=l$, 有

$$v_j^T = Nf_j^*, \quad j = 0, \cdots, l.$$

现在我们考察

$$v_{j+N}^T = \sum_{n=0}^{N-1} f(x_n) e^{-i(j+N)x_n} = \sum_{n=0}^{N-1} f(x_n) e^{-ijx_n},$$

当 $j+N = 0, \cdots, N-1$ 或 $j=-1, \cdots, -N$, 又一次我们有一个吻合,这次是当 $j=-1$ 到 $j=-l$ 时

$$v_{j+N} = Nf_j^*, \quad j = -l, \cdots, -1.$$

不管 $1/N$ 的这个因子,分量 v_i^T 确与系数 f_j^* 吻合,虽然以一种略微混乱一些的顺序. v_j^T 取它们的自然顺序从 v_0^T 到 v_{2l}^T, 容易证明三角函数和系数的顺序为

$$f_0^*, \cdots, f_l^*, \quad f_{-l}^*, \cdots, f_{-1}^*.$$

24.30 对简单的例子 $V=(1,0,-1)$

解 ☞ 这里 $N=3$ 而 $l=1$, 将前题的细节通盘执行一次.

$$3f_j^* = \sum_{n=0}^{2} f(x_n) e^{-ijx_n} = \sum_{n=0}^{2} f(x_n)\omega_3^{jn} = 1 - \omega_3^{2j},$$

这使得 $3f_{-1}^* = 1-\omega_3$, $\quad 3f_0^* = 0$, $\quad 3f_1^* = 1-\omega_3^2$, 于是我们直接地得到三个系数. 转向变换,

$$v_j^T = \sum_{n=0}^{2} f(x_n)\omega_3^{jn} = 1 - \omega_3^{2j}.$$

我们得到

$$v_0^T = 0, \quad v_1^T = 1-\omega_3^2, \quad v_2^T = 1-\omega_3.$$

在题 24.29 中发现的对应被确认.

24.31 在快速 Fourier 变换背后的中心思想是什么?

解 ☞ 当 N 是整数的乘积,数 f_j^* 证明为紧密地互相依赖的. 这种互相依赖性可以被利用来大大地减少生成这些数所需的计算量.

24.32 对最简单的情况,当 N 为二个整数 t_1 及 t_2 的乘积时,推广 FFT.

解 ☞ 令 $j=j_1+t_1j_2$ 和 $n=n_2+t_2n_1$.然后对 $j_1,n_1=0$ 到 t_1-1 和 $j_2,n_2=0$ 到 t_2-1,j 与 n 均按他们所要求的范围从 0 到 $N-1$ 执行.现在

$$\omega^{(j_1+t_1j_2)(n_2+t_2n_1)}=\omega_N^{j_1t_2n_1+j_1n_2+t_1j_2n_2},$$

由于 $t_1t_2=N$ 而且 $\omega_N^N=1$.这个变换可以改写为一个双重和

$$v_j^T=\sum_{n_2=0}^{t_2-1}\sum_{n_1=0}^{t_1-1}v_n\omega_N^{j_1t_2n_1}\omega_N^{j_1n_2+t_1j_2n_2},$$

它也可以安排成一个两-步的算法

$$F_1(j_1,n_2)=\sum_{n_1=0}^{t_1-1}v_n\omega_N^{j_1t_2n_1},$$

$$v_j^T=F_2(j_1,j_2)=\sum_{n_2=0}^{t_2-1}F_1(j_1,n_2)\omega_N^{j_1n_2+t_1j_2n_2}.$$

24.33 假如使用了题 24.32 的 FFT,试问在计算效益中的获利是什么? 换言之快速 Fourier 变换真有那么快吗?

解 ☞ 计算 F_1 要进行 t_1 项,计算 F_2 项有 t_2 项.总共是 t_1+t_2 项.对每个(j_1,n_2)及(j_1,j_2)对,或者说 N 对都必须这么做,因此最终要进行 $N(t_1+t_2)$项的计算,变换最初的形式为

$$v_j^T=\sum_{n=0}^{N-1}v_n\omega_N^{jn},$$

对每个 j 要进行 N 项.总计为 N^2 项,假如按这个标准来度量,效率的收益因而为

$$\frac{t_1+t_2}{N},$$

严重地依赖 N,对一个小的数据集合,譬如说 $N=12=3\times4$,FFT 所需的计算时间为直接处理的 $\frac{7}{12}$,这简直没有多大意义然而它指明的却是事物的走向.

24.34 对下面的向量执行题 24.32 的 FFT:

n	0	1	2	3	4	5		
v_n	0	1	1	0	-1	-1	0	…

解 ☞ 这是一个小尺度的问题,$N=6$,容易看到它的细节,这里 $N=t_1t_2=2\times3$,所以我们首先从

$$F_1(j_1,n_2)=\sum_{n_1=0}^{1}v_n\omega_6^{3j_1n_1},\qquad n=n_2+3n_1$$

求 F_1,取 $\omega=\omega_6$,它们证明如下:

$$F_1(0,0)=v_0+v_3=0,\qquad F_1(1,0)=v_0-v_3=0,$$
$$F_1(0,1)=v_1+v_4=0,\qquad F_1(1,1)=v_1-v_4=2,$$
$$F_1(0,2)=v_2+v_5=0,\qquad F_1(1,2)=v_2-v_5=2.$$

接着

$$v_j^T=F_2(j_1,j_2)=\sum_{n_2=0}^{2}F_1(j_1,n_2)\omega^{j_1n_2+t_1j_2n_2}.$$

由于 $j=j_1+2j_2$ 导出,

$$v_0^T=F_2(0,0)=v_0+v_1+v_2+v_3+v_4+v_5=0,$$
$$v_1^T=F_2(1,0)=F_1(1,0)+F_1(1,1)\omega+F_1(1,2)\omega^2=2\omega+2\omega^2=2\sqrt{3}i,$$
$$v_2^T=F_2(0,1)=F_1(0,0)+F_1(0,1)\omega^2+F_2(0,2)\omega^4=0,$$
$$v_3^T=F_2(1,1)=F_1(1,0)+F_1(1,1)\omega^3+F_1(1,2)\omega^6=0.$$

类似地

$$v_4^T = F_2(0,2) = 0,$$
$$v_5^T = F_2(1,2) = -2\sqrt{3}i.$$

注意到在计算 F_1 值时包含了 Nt_1 项在得到 F_2 时包含了 Nt_2 个项,总共 12+18=30 项,直接计算要用 36 项,因而确认刚才所得到的结果,还要注意到 j_1,j_2 进行的顺序,在程序语言中 j_2 对 j_1 循环来说是外循环.

24.35　将题 24.32 的 FFT 推广到 $N = t_1t_2t_3$ 的情况.

解　☞ 细节将提示推广到还要更长一些乘积的方法.令

$$j = j_1 + t_1j_2 + t_1t_2j_3, \qquad n = n_3 + t_3n_2 + t_3t_2n_1,$$

并且考察在

$$\omega_N^{(j_1+t_1j_2+t_1t_2j_3)(n_3+t_3n_2+t_3t_2n_1)}$$

中的 9 个可能的幂项,三个含有乘积 $t_1t_2t_3$ 的项可以被忽略掉,由于 $\omega_N^N = 1$.剩下的 6 个项可以在变换中作如下分组,

$$v_j^T = \sum_{n_3=0}^{t_3-1}\left[\sum_{n_2=0}^{t_2-1}\left(\sum_{n_1=0}^{t_1-1} v_n\omega_N^{j_1t_3t_2n_1}\right)\omega_N^{(j_1+t_1j_2)t_3n_2}\right]\omega_N^{(j_1+t_1j_2+t_1t_2j_3)n_3}$$

n_1 只出现在内层和中而 n_2 不出现在外层和中,如前,这个三重和可以表示成一个算法,现在分成三步走

$$F_1(j_1,n_2,n_3) = \sum_{n_1=0}^{t_1-1} v_n\omega_N^{j_1t_3t_2n_1},$$

$$F_2(j_1,j_2,n_3) = \sum_{n_2=0}^{t_2-1} F_1(j_1,n_2,n_3)\omega_N^{(j_1+t_1j_2)t_3n_2},$$

$$v_j^T = F_3(j_1,j_2,j_3) = \sum_{n_3=0}^{t_3-1} F_2(j_1,j_2,n_3)\omega_N^{(j_1+t_1j_2+t_1t_2j_3)n_3},$$

这就是所要求的 FFT.

24.36　估计假如使用这个算法所节约的计算时间.

解　☞ 在三步中的每一步三元组的数目,诸如 (j_1,n_2,n_3) 必须要进行的是 $t_1t_2t_3 = N$.在这个和中我们发现项的数目依次为 t_1,t_2,t_3,这就造成了总数合起来为 $N(t_1+t_2+t_3)$ 项.该变换正如所定义的那样还是用了 N^2 项,所以 FFT 的效率可以估作

$$\frac{t_1+t_2+t_3}{N}.$$

例如,$N = 1000 = 10\times10\times10$,则只是原来所需 1,000,000 项的百分之 3.

24.37　对输入向量

n	0	1	2	3	4	5	6	7
v_n	1	$1+i$	i	$i-1$	-1	$-1-i$	$-i$	$1-i$

以手工方法执行题 24.35 中的 FFT 算法.

解　☞ 我们有 $N = 8 = 2\times2\times2$,使 $j = j_1+2j_2+4j_3$ 与 $n = n_3+2n_2+4n_1$.于是关于 F_1 的公式为

$$F_1(j_1,n_2,n_3) = \sum_{n_1=0}^{1} v_n\omega_8^{4j_1n_1},$$

以及我们有

$$F_1(0,0,0)=v_0+v_4=0,\quad F_1(1,0,0)=v_0+v_4\omega^4=2,$$
$$F_1(0,0,1)=v_1+v_5=0,\quad F_1(1,0,1)=v_1+v_5\omega^4=2+2i,$$
$$F_1(0,1,0)=v_2+v_6=0,\quad F_1(1,1,0)=v_2+v_6\omega^4=2i,$$
$$F_1(0,1,1)=v_3+v_7=0,\quad F_1(1,1,1)=v_3+v_7\omega^4=2i-2,$$

将 ω_8 缩写成 ω.注意用了 $Nt_1=8\times 2$ 项,接下来我们用

$$F_2(j_1,j_2,n_3)=\sum_{n_2=0}^{1}F_1(j_1,n_2,n_3)\omega^{2(j_1-2j_2)n_2}$$

来计算

$$F_2(0,0,0)=0,\quad F_2(1,0,0)=F_1(1,0,0)+F_1(1,1,0)\omega^2=4,$$
$$F_2(0,0,1)=0,\quad F_2(1,0,1)=F_1(1,0,1)+F_1(1,1,1)\omega^2=4+4i,$$
$$F_2(0,1,0)=0,\quad F_2(1,1,0)=F_1(1,0,0)+F_1(1,1,0)\omega^6=0,$$
$$F_2(0,1,1)=0,\quad F_2(1,1,1)=F_1(1,0,1)+F_1(1,1,1)\omega^6=0,$$

以及最后有

$$v_j^T=F_3(j_1,j_2,j_3)=\sum_{n_3=0}^{1}F_2(j_1,j_2,n_3)\omega^{jn_3}.$$

便得到变换

$$v_0^T=F_3(0,0,0)=F_2(0,0,0)+F_2(0,0,1)=0,$$
$$v_1^T=F_3(1,0,0)=F_2(1,0,0)+F_2(1,0,1)\omega=4+4\sqrt{2},$$
$$v_2^T=F_3(0,1,0)=F_2(0,1,0)+F_2(0,1,1)\omega^2=0,$$
$$v_3^T=F_3(1,1,0)=F_2(1,1,0)+F_2(1,1,1)\omega^3=0,$$
$$v_4^T=F_3(0,0,1)=F_2(0,0,0)+F_2(0,0,1)\omega^4=0,$$
$$v_5^T=F_3(1,0,1)=F_2(1,0,0)+F_2(1,0,1)\omega^5=4-4\sqrt{2},$$
$$v_6^T=F_3(0,1,1)=F_2(0,1,0)+F_2(0,1,1)\omega^6=0,$$
$$v_7^T=F_3(1,1,1)=F_2(1,1,0)+F_2(1,1,1)\omega^7=0.$$

进行了总和为 $N(t_1+t_2+t_3)=48$ 项的运算从,$N^2=64$ 中只有少量的节约,其原因是问题是小尺度的.

24.38 逆离散变换可以定义为

$$u_k^{-T}=\frac{1}{N}\sum_{j=0}^{N-1}u_j\omega^{-jk}=\frac{1}{N}\sum_{j=0}^{N-1}u_je^{ikx_j},$$

通过插入 $u_j=v_j^T$ 来证明这个定义确给出一个逆关系并发现它就是 $u_k^{-T}=v_k$ 也就是说再一次获得原始向量的分量.

解 ☞ 首先我们利用关系式

$$\omega^{jn}=e^{-ijx_n}$$

将题24.31的结果加以改写,对区间$(0,N-1)$中的 j,k 来得到

$$\sum_{n=0}^{N-1}\omega^{jn}\omega^{-kn}=\begin{cases}N, & 若\ k=j,\\ 0, & 若\ k\neq j,\end{cases}$$

或许是有用的.现在

$$\frac{1}{N}\sum_{j=0}^{N-1}v_j^T\omega^{-jk}=\frac{1}{N}\sum_{j=0}^{N-1}\sum_{n=0}^{N-1}v_n\omega^{jn}\omega^{-jk}=\frac{1}{N}\sum_{n=0}^{N-1}v_n\sum_{j=0}^{N-1}\omega^{(n-k)j}$$

最后的和为零,除非将 n 取作 k.我们立得所期望的 v_k.

24.39 求题24.37中获得的变换的逆变换.

解 ☞ 可以使用FFT,然而考虑到分量大量为0,这是一个直接进行的好机会

$$8u_0^{-T}=\sum_{j=0}^{7}v_j^T=8,\qquad u_0^{-T}=1=v_0,$$

$$8u_1^{-T}=\sum_{j=0}^{7}v_j^T\omega^{-j}=(4+4\sqrt{2})\omega^{-1}+(4-4\sqrt{2})\omega^{-5}$$

$$= 8(1+i), \qquad u_1^{\ T} = 1 + i = v_2,$$

$$8u_2^{\ T} = \sum_{j=0}^{7} v_j^T \omega^{-2j} = (4+4\sqrt{2})\omega^{-2} + (4-4\sqrt{2})\omega^{-10}$$

$$= 8i, \qquad u_2^{\ T} = i = v_3,$$

$$\vdots$$

余下的分量可以仿照题 24.63 那样加以证明.

补 充 题

24.40 应用题 24.2 的方法于下面的数据.

x	0	1	2	3	4
y	0	1	2	1	0

24.41 导出题 24.5 的系数公式.

24.42 应用题 24.5 的方法于下面的数据.

x	0	1	2	3	4	5
y	0	1	2	2	1	0

24.43 使用题 24.6 的结果对题 24.40 的数据来求得最小二乘方和 $T_0(x)$ 及 $T_1(x)$.

24.44 仿效题 24.6 的论点去获得一个关于偶数个 x 自变量的多少有点类似的结果.

24.45 将前题的结果应用于题 24.42 的数据.

24.46 推广题 24.40 的数据于一个周期为 8 的奇函数,求得一个正弦函数的和来表示这个函数.

24.47 推广题 24.40 的数据于一个周期为 8 的偶函数,求得一个余弦函数的和来表示这个函数.

24.48 证明对 $y(x)=|\sin x|$ 的 Fourier 级数的"完全地纠正的"(fully rectified)正弦波为

$$y(x) = \frac{4}{\pi}\left(\frac{1}{2} - \frac{\cos 2x}{1\cdot 3} - \frac{\cos 4x}{3\cdot 5} - \frac{\cos 6x}{5\cdot 7} - \cdots\right).$$

24.49 证明关于 $y(x)=x^2$ 当 x 在 $-\pi$ 与 π 之间时具有 2π 周期的 Fourier 级数为

$$y(x) = \frac{\pi^2}{3} - 4\sum_{k=1}^{\infty}\frac{(-1)^{k-1}\cos kx}{k^2}.$$

使用这个结果计算级数 $\sum_{k=1}^{\infty}(-1)^{k-1}/k^2$ 及 $\sum_{k=1}^{\infty}1/k^2$.

24.50 使用题 24.15 的 Fourier 级数计算 $\sum_{k=1}^{\infty}1/(2k-1)^2$.

24.51 使用题 24.16 的 Fourier 级数去证明 $\pi/4 = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$.

24.52 使用题 24.17 的级数去计算 $1 - 1/3^3 + 1/5^3 - 1/7^3 + \cdots$.

24.53 什么是对题 24.48 中的函数的 4-项的最小二乘方三角逼近? 什么是 2-项的最小二乘方逼近?

24.54 应用 Fourier 平滑技术于下面的数据,假设端点值是真正为零的并延拓该函数为一个奇函数.还要试验其他的平滑方法,或是组合的平滑法.与准确值 $y(x)=x(1-x)$ 比较结果,给定的数据是来自于它再加上一个直至 20% 的随机量获得的,自变量为 $x=0(0.05)1$.这些数据是 0.00, 0.06, 0.10, 0.11, 0.14, 0.22, 0.22, 0.27, 0.28, 0.21, 0.22, 0.27, 0.21, 0.20, 0.19, 0.21, 0.19, 0.12, 0.08, 0.04, 0.00.

24.55 证明在引言节中给出的系数关系

$$a_j = c_j + c_{-j}, \qquad b_j = i(c_j - c_{-j})$$

以及逆关系

$$c_j = \frac{a_j - ib_j}{2}, \qquad c_{-j} = \frac{a_j + ib_j}{2}.$$

推导若 a_j, b_j 为实时，则 c_j 及 c_{-j} 必定为共轭复数. 回忆对配置三角多项式，我们有 $c_j = f_j^*$，并假设 a_j, b_j 及 $f(x)$ 均为实数，证明

$$a_j = 2\mathrm{Re}(f_j^*) = \frac{2}{N}\sum_{n=0}^{N-1} f(x_n)\cos jx_n,$$

$$b_j = -2\mathrm{Im}(f_j^*) = \frac{2}{N}\sum_{n=0}^{N-1} f(x_n)\sin jx_n.$$

24.56 如在题 24.30 中那样进行，使用 $V=(1,-1,0)$.

24.57 如在题 24.34 中那样进行，使用的 V 向量为：

n	0	1	2	3	4	5
v_n	0	0	1	1	1	0

24.58 如在题 24.37 中那样进行使用的 V 向量为：

n	0	1	2	3	4	5	6	7
v_n	1	$1+i$	0	$1-i$	0	$1+i$	0	$1-i$

24.59 通过应用原始变换

$$v_j^T = \sum_{n=0}^{N-1} v_n \omega_N^{jn}$$

来确认题 24.58 的结果.

24.60 使用初等微积分证明若 $t_1 t_2 = N$，则 $t_1 + t_2$ 的极小值当 $t_1 = t_2$ 时出现. 将这个结果推广到 $t_1 t_2 t_3 = N$ 的情况. 对 FFT 来说它隐含的是什么？

24.61 求在题 24.30 中获得的变换的逆变换.

24.62 应用题 24.32 中的 FFT 使题 24.34 的输出数据倒转.

n	0	1	2	3	4	5
v_n^T	0	$2\sqrt{3}i$	0	0	0	$-2\sqrt{3}i$

24.63 完成题 24.39 中开始的逆转.

24.64 使用 FFT 做同样的逆转.

第二十五章　非线性代数

方程的根

本章处理的是方程的求根，或是方程组的求根的占老问题．有用方法的长长列表说明该问题的悠久历史以及它的持续重要性，使用哪一个方法则依赖于人们是否要求一个特殊问题的所有的根或是仅仅是其中的几个，依赖于根是实的还是复的，单根还是重根，人们有没有一个现成的首次逼近，等等．

1．**迭代法**解 $x=F(x)$ 通过一个递推公式

$$x_n = F(x_{n-1}).$$

若 $|F'(x)|\leqslant L<1$ 则 x_n 收敛于一个根．误差为 $e_n=r-x_n$，其中 r 是精确解，误差具有性质

$$e_n \approx F'(r)e_{n-1},$$

所以每迭代一次便将误差减少一个近乎 $F'(r)$ 的因子．若 $F'(r)$ 接近 1 的话，这个收敛就是慢的．

2．在有些情况下 **Δ^2 过程可以加速收敛**，它含有一个逼近公式

$$r \approx x_{n+2} - \frac{(\Delta x_{n+1})^2}{\Delta^2 x_n},$$

它可以由上面给出的误差性质导出．

3．**Newton 方法**获得对 $f(x)=0$ 的一个根的逐次逼近

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})},$$

因而毫无疑问地是一个十分受欢迎的算法，在 $f'(x)$ 是复杂的情况下，前述迭代法可能是可取的，然而 Newton 法收敛要快得多通常得到首肯．这里误差满足

$$e_n \approx -\frac{f''(r)}{2f'(r)}e_{n-1}^2,$$

称它为**平方收敛**，每一次误差粗略地与前一次误差的平方成正比，每次迭代准确位数几乎增加一倍．

平方根迭代

$$x_n = \frac{1}{2}\left(x_{n-1} + \frac{Q}{x_{n-1}}\right)$$

是 Newton 法的一种特殊情况，它相应于取 $f(x)=x^2-Q$．当 $Q>0$ 时它平方地收敛于 Q 的正平方根．

更一般的寻找根的公式为

$$x_n = x_{n-1} - \frac{x_{n-1}^p - Q}{px_{n-1}^{p-1}},$$

它也是 Newton 法的一种特殊情况，它产生 Q 的一个 p 次根．

4．**插值法**用二次或更多次的逼近，通常某些太小而某些又太大，为了得到一个改进的近似根可以使用配置多项式．这些方法中最古老的就是建立在前二次逼近的线性插值基础上的．它被称作试位法(regula falsi)，它是用迭代公式

$$x_n = x_{n-1} - \frac{(x_{n-1}-x_{n-2})f(x_{n-1})}{f(x_{n-1})-f(x_{n-2})}$$

来解 $f(x)=0$ 的，收敛速度介于前二个方法之间，一种基于在前三个逼近值 x_0,x_1,x_2

之间的二次插值，使用的公式为

$$x_3 = x_2 - \frac{2C}{B \pm \sqrt{B^2 - 4AC}}.$$

关于 A, B, C 的表达式将在题 25.8 中给出.

5. **Bernoulli** 方法产生一个实多项式

$$a_0x^n + a_1x^{n-1} + \cdots + a_n = 0$$

的最大根，假如存在一个单的最大根的话，通过计算差分方程的一个解序列

$$a_0x_k + a_1x_{k-1} + \cdots + a_nx_{k-n} = 0,$$

并且取极限 $\lim(x_{k+1}/x_k)$. 通常使用初始值 $x_{-n+1} = \cdots = x_{-1} = 0, x_0 = 1$, 如果一个复共轭对为最大根的话，那么解序列仍可计算下去，但是用来决定根，$r_1, r_2 \approx r(\cos\phi \pm i\sin\phi)$的公式是

$$r^2 \approx \frac{x_k^2 - x_{k+1}x_{k-1}}{x_{k-1}^2 - x_kx_{k-2}}, \qquad -2r\cos\phi \approx \frac{x_{k+1}x_{k-2} - x_{k-1}x_k}{x_{k-1}^2 - x_kx_{k-2}}.$$

6. **降阶法**(Deflation)指的是从多项式方程中移出一个已知根的过程，从而导出一个较低次的新方程，它与 Bernoulli 方法联在一起，便允许一个接一个地发现下一个主根. 在实践中发现以持续降阶去决定较小的根精度会减小. 然而，使用这些所得到的结果，在每一步上可以作为 Newton 方法的初始逼近值往往带来所有根的精确计算.

7. **商-差算法**(quotient-difference)推广 Bernoulli 的方法，它可以同时地产生一个多项式方程包括复共轭对的所有的根. 它包含计算一个商和差分的表(类似于差分表). 从这个表中推导出根，其细节多少有点复杂，可在题 25.25 到题 25.32 中找到.

8. **Sturm 序列**提供对一个方程的实根的另一个历史性的逼近还是基本上同时地产生它们. 一个 Sturm 序列

$$f_0(x), f_1(x), \cdots, f_n(x)$$

满足题 25.33 中所列出的 5 个条件，这些条件保证 $f_0(x)$在区间(a,b)中的实根个数精确地为序列 $f_0(a), f_1(a), \cdots, f_n(a)$的变号数与序列 $f_0(b), f_1(b), \cdots, f_n(b)$的变号数之差，通过选择不同的区间$(a,b)$，因而实根的位置可以被确定，当 $f_0(x)$是一个多项式，一个适当的 Sturm 序列通过用欧几里得(Euclide 算法)求得，令 $f_1(x) = f'_0(x)$序列的余下部分定义为

$$f_0(x) = f_1(x)L_1(x) - f_2(x),$$
$$f_1(x) = f_2(x)L_2(x) - f_2(x),$$
$$\cdots\cdots$$
$$f_{n-2}(x) = f_{n-1}(x)L_{n-1}(x) - f_n(x).$$

像降阶法与商-差分法，Sturm 序列可以用来获得 Newton 法的好的出发近似值，然后 Newton 法再以高速产生精确度高的根.

方程组与最优化问题

由前面许多方法的推广，以及其他的算法也适用于方程组，我们选择其中的三个方法

1. **迭代法**，例如，以公式

$$x_n = F(x_{n-1}, y_{n-1}), \quad y_n = G(x_{n-1}, y_{n-1})$$

解一对方程

$$x = F(x,y), \qquad y = G(x,y),$$

假设 x_n 与 y_n 二个序列均收敛. Newton 法解

$$f(x,y) = 0, \qquad g(x,y) = 0.$$

通过由

$$x_n = x_{n-1} + h_{n-1}, \quad y_n = y_{n-1} + k_{n-1}$$

所定义的序列,其中的 h_{n-1} 及 k_{n-1},由

$$f_x(x_{n-1}, y_{n-1})h_{n-1} + f_y(x_{n-1}, y_{n-1})k_{n-1} = -f(x_{n-1}, y_{n-1}),$$
$$g_x(x_{n-1}, y_{n-1})h_{n-1} + g_y(x_{n-1}, y_{n-1})k_{n-1} = -g(x_{n-1}, y_{n-1})$$

所确定.

更一般地,对于方程组

$$F(x) = 0,$$

其中,F, x 及 0 为 n 维向量,迭代法

$$x^{(n)} = G(x^{(n-1)})$$

也适用于它,它通过对原始方程组的一个重新安排,加上一个适当的初始向量 $x^{(0)}$ 而得到.或是 Newton 方法可以用一个紧凑的向量-矩阵形式来表示,从 Taylor 级数开始

$$F(x^{(n-1)} + h) = F(x^{(n-1)}) + J(x^{(n-1)})h + \cdots,$$

忽略更高阶的项并令左侧为零向量,其结果是一个关于 h 的线性方程组

$$J(x^{(n-1)})h = -F(x^{(n-1)}),$$

甚至可以将它写成

$$h = -J^{-1}(x^{(n-1)})F(x^{(n-1)}),$$

J 称作 F 的 Jacobian 矩阵,它具有元素

$$J_{ij} = \frac{\partial f_i}{\partial x_j},$$

其中 f_i 与 x_j 为 F 与 x 的分量.用一个精确的初始逼近值,和一个适当的 F,误差在

$$\| x - x^{(n)} \| \leqslant c \| x - x^{(n-1)} \|^2$$

的意义下下降,称作平方地下降,但是必须指出这个平方收敛可以是难以捉摸的.由于寻找充分精确的初始逼近值对方程组而言不总是那么容易的,因而 Newton 逼近有时会离开正道.在有些情况下发现简化的步骤

$$x^{(n)} = x^{(n-1)} + kn, \qquad k < 1$$

工作得更好,选择 k 保证 F 的范数的下降.以这种方法在每一步上情况都有改进.这种方法称为阻尼的 Newton 法.

2. **最优化方法**是基于这样的思想,即方程组 $F=0$ 或 $f_i=0, i=1,2,\cdots,n$,当函数

$$S = f_1^2 + f_2^2 + \cdots + f_n^2$$

被极小化时,求得它的解.因为极小明显地出现在当所有 f_i 为零时.方程组就算被解了,寻找这个极小值的直接法或是下降法都是曾经被推出过的.例如,二维问题(用一个熟悉的记号变换)

$$f(x, y) = 0, \qquad g(x, y) = 0$$

等价于将和

$$S(x, y) = f^2 + g^2$$

极小化.从一个初始逼近值(x_0, y_0)开始,我们选择下一个逼近值其形式为:$x_1 = x_0 - tS_{x0}$, $y_1 = y_0 - tS_{y0}$,其中 S_{x0}及 S_{y0}为 S 在(x_0, y_0)处梯度向量的分量.因此,按最速下降的方向前进,而该算法称作最速下降算法,t 这个数可以选作在这个方向上将 S 极小化的,虽然曾经提出过各种方法,然而接着下来的是类似的步骤.这个方法常被用作对 Newton 法提供初始逼近.

当然,与上面等价的往往是按相反方向来开发的.优化一个函数 $f(x_1, \cdots, x_n)$ 人们寻找 f 的梯度为零的地方

$$\text{grad}(f) = (f_1, f_2, \cdots, f_n) = (0, 0, \cdots, 0),$$

此处 f_i 表示 f 对 x_i 的偏导数,最优化就是尝试对这 n 个非线性方程的方程组求解.

3. **Bairstow 方法**是通过应用 Newton 法于一个相关的方程组来产生一个实多项式方程 $p(x)=0$的复根.更明确地,以一个二次多项式除 $p(x)$,得出一个恒等式

$$p(x) = (x^2 - ux - v)q(x) + r(x),$$

其中 $r(x)$为一个线性余项

$$r(x) = b_{n-1}(u,v)(x-u) + bn(u,v),$$

这个二次的除数将是 $p(x)$的一个因子,如果我们能选择 u,v 满足

$$b_{n-1}(u,v) = 0, \quad b_n(u,v) = 0$$

的话,将 Newton 法应用在这个方程组上,一旦 u,v 为已知的,一对复根可以通过解

$$x^2 - ux - v = 0$$

求得.

题　解

迭代方法

25.1 证明若 r 为$f(x)=0$ 的一个根且若该方程改写为形式 $x=F(x)$,而且 $F(x)$在一个以 $x=r$ 为中心的区间I 中满足$|F'(x)|\leqslant L<1$,取 x_0 为任意的,但是在区间 I 中,则序列 $x_n=F(x_n-1)$,有极限 $\lim x_n=r$.

证 ☞ 首先我们得到

$$|F(x)-F(y)|=|F'(\xi)(x-y)|\leqslant L|x-y|.$$

假设 x,y 都接近r,事实上这只是 Lips chitz 条件.而不需在我们所要的 $F'(x)$上加上更严格的限制条件.

$$|x_n-r|=|F(x_{n-1})-F(r)|\leqslant L|x_{n-1}-r|,$$

于是,由于 $L<1$,每一个逼近至少像它的前任那么好.这就保证了我们所有的逼近均落在 I 中,因而没有什么会使程序中断.应用最后一个不等式 n 次,我们有

$$|x_n-r|\leqslant L^n|x_0-r|$$

以及由于 $L<1$,$\lim x_n=r$ 成立.

这个收敛在图 25.1 中得到说明.注意选择 $F(x_{n-1})$作为下一个 x_n,相当于跟踪一个水平直线到 $y=x$ 线*.同时注意在图 25.2 中$|F'(x)|>1$ 的情况导致发散.

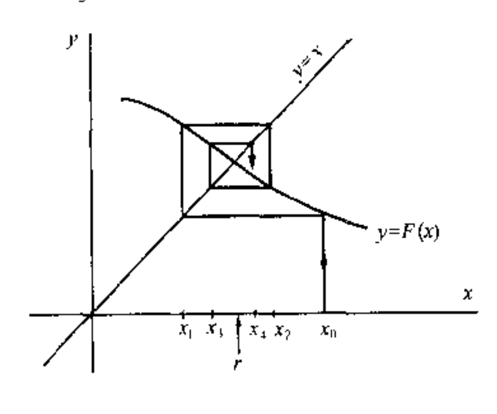


图 25.1

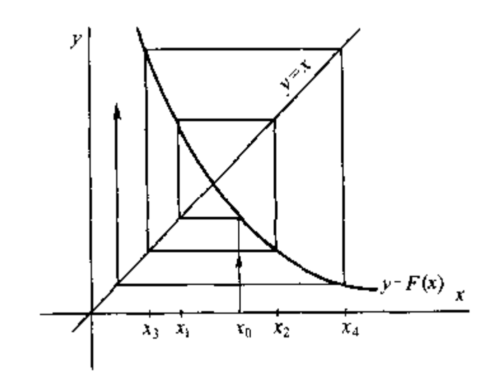


图 25.2

25.2 在 1225 年 Pisa 的 Leonardo 研究了方程

$$f(x) = x^3 + 2x^2 + 10x - 20 = 0,$$

* 译注:数学家华罗庚形象地称这种方法为螺丝转方法.并可推广到二根任意的非线性曲线的求解上(即求它们的公共零点),任一个初学者均可从这一思想中得到启示.

并且得到 $x=1.368,808,107$. 没有人知道 Leonardo 用什么方法得到这个值的,然而在他的时代这是一个值得注意的结果.应用题 25.1 的方法来得到这个结果.

解 ☞ 这个方程可以用多种方法改写成 $x=F(x)$ 的形式,我们取 $x=F(x)=20/(x^3-2x+10)$ 它给出了迭代公式

$$x_n=\frac{20}{x_{n-1}^2+2x_{n-1}+10},$$

取 $x_0=1$ 我们得到 $x_1=\frac{20}{13}\approx1.538461538$,继续这一迭代产生表 25.1 的序列.肯定 24 次迭代就足够了,那时 Leonardo 值出现.

表 25.1

n	x_n	n	x_n
1	1.538461538	13	1.368817874
2	1.295019157	14	1.368803773
3	1.401825309	15	1.368810031
4	1.354209390	16	1.368807254
5	1.375298092	17	1.368808486
6	1.365929788	18	1.368807940
7	1.370086003	19	1.368808181
8	1.368241023	20	1.368808075
9	1.369059812	21	1.368808122
10	1.368696397	22	1.368808101
11	1.368857688	23	1.368808110
12	1.368786102	24	1.368808107

25.3 为什么前题中的算法收敛得如此地慢?

解 ☞ 收敛速度可以从关系

$$e_n=r-x_n=F(r)-F(x_{n-1})=F'(\xi)(r-x_{n-1})=F'(\xi)e_{n-1}$$

估得,它将 n 次误差 e_n 与前一个误差 e_{n-1} 相比较.当 n 增加时我们可以取 $F'(r)$ 为 $F'(\xi)$ 的一个逼近,假设这个导数是存在的话,于是 $e_n\approx F'(r)e_{n-1}$.在我们的例子中,

$$F'(r)=-\frac{40(r+1)}{(r^2+2r+10)^2}\approx-0.44$$

使每一个误差约为前一个的 -0.44 倍.它提示了对每一个新的准确小数位要求 2—3 次迭代,而这就是算法实际能完成的.

25.4 应用外推到极限的思想来加速前面的算法.

解 ☞ 当在一个算法中有误差特征的信息可用时,这个思想可以被使用.这里我们有逼近值 $e_n\approx F'(r)e_{n-1}$.无需 $F'(r)$ 的知识我们还是可以写成

$$r-x_{n+1}\approx F'(r)(r-x_n),$$
$$r-x_{n+2}\approx F'(r)(r-x_{n+1}).$$

将它们相除我们得到

$$\frac{r-x_{n+1}}{r-x_{n+2}}\approx\frac{r-x_n}{r-x_{n+1}},$$

并将根解出

$$r\approx x_{n+2}-\frac{(x_{n+2}-x_{n+1})^2}{x_{n+2}-2x_{n+1}+x_n}$$
$$=x_{n+2}-\frac{(\Delta x_{n+1})^2}{\Delta^2x_n},$$

这就是常被称作的 **Aitken Δ^2 过程**.

25.5 对题 25.2 的计算应用外推到极限法.

解 ☞ 使用 x_{10}, x_{11} 及 x_{12} 由公式可得

$$r \approx 1.368786102 - \frac{(0.000071586)^2}{-0.000232877}$$
$$\approx 1.368808107.$$

它又一次为 Leonardo 值,用这种外推,只需一半的迭代.早一些使用它,可以促成收敛方面的更进一步的节约.

25.6 在每三次迭代后对称地使用外推到极限法就是那个称作 Steffensen 的方法.将它应用于 Leonardo 方程.

解 ☞ 头三个逼近值 x_0, x_1 及 x_2 可以向题 25.2 借用,现在用 Aitken 的公式来产生 x_3:

$$x_3 = x_2 - \frac{(x_2 - x_1)^2}{x_2 - 2x_1 + x_0} = 1.370813882.$$

原始迭代现在被恢复成像在题 25.2 中那样来得到 x_4 及 x_5:

$$x_4 = F(x_3) = 1.367918090, \quad x_5 = F(x_4) = 1.369203162.$$

Aitken 公式接着提供 x_6:

$$x_6 = x_5 - \frac{(x_5 - x_4)^2}{x_5 - 2x_4 + x_3} = 1.368808169.$$

下一个循环带来的迭代值为

$$x_7 = 1.368808080, \quad x_8 = 1.368808120.$$

Aitken 公式从这些值得出 $x_9 = 1.368808108$.

25.7 证明将 Leonardo 方程作其他的重新安排不一定产生收敛序列.

证 ☞ 作为例子我们可以取 $x = (20 - 2x^2 - x^3)/10$ 由它得出的迭代公式为

$$x_n = \frac{20 - 2x_{n-1}^2 - x_{n-1}^3}{10},$$

再一次以 $x_0 = 1$ 为出发值,我们被带往序列

$$x_1 \approx 1.70 \quad x_3 \approx 1.75 \quad x_5 \approx 1.79 \quad x_7 = 1.83$$
$$x_2 \approx 0.93 \quad x_4 \approx 0.85 \quad x_6 \approx 0.79 \quad x_8 \approx 0.72$$

等等.看来是清楚的交替的逼近朝着相反的方向,与题 25.1 相比较我们发现这里 $F'(r) = (-4r - 3r^2)/10 < -1$,它确认了计算的根据.

Newton 法

25.8 导出解 $f(r) = 0$ 的 Newton 迭代公式

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}.$$

解 ☞ 从 Taylor 公式开始

$$f(r) = f(x_{n-1}) + (r - x_{n-1})f'(x_{n-1}) + \frac{1}{2}(r - x_{n-1})^2 f''(\xi).$$

我们保留线性部分,回忆 $f(r) = 0$,并通过将 x_n 去取代还剩留在公式中的 r 去得到

$$0 = f(x_{n-1}) + (x_n - x_{n-1})f'(x_{n-1}).$$

将它重新安排立得 $r \approx x_n = x_{n-1} - \dfrac{f(x_{n-1})}{f'(x_{n-1})}$.

25.9 Newton 公式的几何解释是什么?

解 ☞ 它相当于使用在 x_{n-1} 处以对 $y = f(x)$ 的切线来代替曲线.在图 25.3 中可以看到它导出

$$\frac{f(x_{n-1}) - 0}{x_{n-1} - x_n} = f'(x_{n-1}).$$

它还是 Newton 公式,接下来的类似步骤如箭头所示

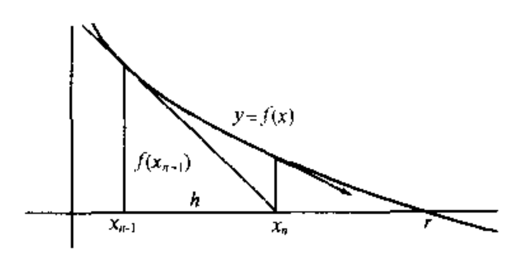


图 25.3

25.10 将 Newton 公式用于 Leonardo 方程.

解 ☞ 取 $f(x)=x^3+2x^2+10x-20$ 我们得到 $f'(x)=3x^2+4x+10$,而迭代公式就变成

$$x_n=x_{n-1}-\frac{x_{n-1}^3+2x_{n-1}^2+10_{n-1}-20}{3x_{n-1}^2+4x_{n-1}+10}.$$

再一次选 $x_0=1$,我们得到表 25.2 的结果,

表 25.2

n	1	2	3	4
x_n	1.411764706	1.369336471	1.368808189	1.368808108

收敛速度是显著的.在四次迭代中基本上得到 Leonardo 值.事实上,计算证明

$$f(1.368808107)\approx-0.000000016,$$

$$f(1.368808108)\approx-0.000000005,$$

它提示了 Newton 结果是略胜一筹.

25.11 通过展示收敛是"二次的"来解释 Newton 迭代的快速收敛.

解 ☞ 回忆题 25.8 的方程,它导出 Newton 公式

$$f(r)=f(x_{n-1})+(r-x_{n-1})f'(x_{n-1})+\frac{1}{2}(r-x_{n-1})^2f''(\xi),$$

$$0=f(x_{n-1})+(x_n-x_{n-1})f'(x_{n-1}),$$

相减得到 $0=(r-x_n)f'(x_{n-1})+\frac{1}{2}(r-x_{n-1})^2f''(\xi)$,或者令 $e_n=r-x_n$,$0=e_nf'(x_{n-1})+\frac{1}{2}e_{n-1}^2f''(\xi)$,假设收敛,我们将 x_{n-1}及 ξ 都换成 r,于是我们有

$$e_n\approx-\frac{f''(r)}{2f'(r)}e_{n-1}^2.$$

因此每一个误差粗略地与前一误差的平方成正比,这意味着每次逼近准确的小数位粗略地增加了一倍,因而为什么称之为平方收敛.它可以与在题 25.3 中的较慢的线性收敛相比,那里每个误差粗略地正比于前一误差.由于我们现在的 x_3 的误差约为 0.00000008,而$[f''(r)]/[2f'(r)]$约为 0.3,我们发现如果在我们的计算中能用更多位小数进行工作,x_4 的误差就大约是在第 15 位上的二个单位!这种超速度提示了 Newton 算法值得启动它作一个合理精确的首次逼近,而它的自然角色是把这样一个合理的逼近转为一个极好的.事实上,其他出现的算法比之 Newton 方法更适合于求所有根的首次逼近值的"全局"问题.然而这类方法通常收敛得很慢.看来只能自然地使用它们仅仅作为一个合理的首次逼近,然后用 Newton 法提供修饰,这类方法很受欢迎因而我们在进行中会再一次提到它,但是偶尔也会发现这样的情况,即给出的是一个不适当的首次逼近.Newton 法将以二次速度收敛,但并不是到所希望的那个根!回忆算法背后的切线几何,容易绘出一条会发生这种情况的曲线,简单地把首次逼近安放在极大或极小点的近旁就是如此.

25.12 证明决定平方根的公式

$$x_n=\frac{1}{2}\left(x_{n-1}+\frac{Q}{x_{n-1}}\right)$$

是 Newton 迭代的一个特殊情况.

证 ☞ 取 $f(x)=x^2-Q$，显然使 $f(x)=0$ 相当于找 Q 的一个平方根，由于 $f'(x)=2x$，Newton 公式变成

$$x_n = x_{n-1} - \frac{x_{n-1}^2 - Q}{2x_{n-1}} = \frac{1}{2}\left(x_{n-1} + \frac{Q}{x_{n-1}}\right).$$

25.13 应用平方根迭代取 $Q=2$.

解 ☞ 选 $x_0=1$，我们得到表 25.3 中的结果，再一次看到收敛的二次性质. 每一个结果粗略地具有二倍于它的前者那么多的准确位数，图 25.4 说明这个行为，由于首次逼近是在 $y=x^2-2$ 的凹的那一侧，下一个则是在根的另一侧. 在这以后序列是单调的. 剩下的在曲线的凸的一侧像切线通常所做的那样.

表 25.3

n	x_n
1	1.5
2	1.416 666 667
3	1.414 215 686
4	1.414 213 562
5	1.414 213 562

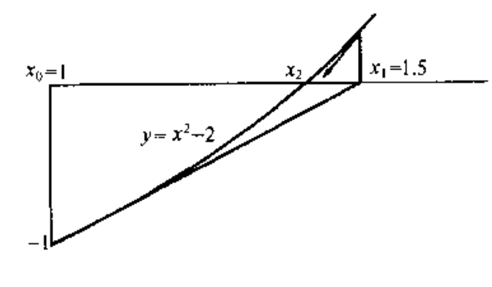


图 25.4

25.14 导出找 Q 的 p 次根的迭代公式 $x_n = x_{n-1} - \dfrac{x_{n-1}^p - Q}{p x_{n-1}^{p-1}}$.

解 ☞ 取 $f(x)=x^p-Q$ 及 $f'(x)=px^{p-1}$，这个结果立刻成为 Newton 法的一个特殊情况.

25.15 应用前题寻找 2 的一个三次根.

解 ☞ 取 $Q=3$ 及 $p=3$，迭代公式简化成 $x_n=\frac{2}{3}(x_{n-1}+\frac{1}{x_{n-1}^2})$ 选择 $x_0=1$，我们得到 $x_1=\frac{4}{3}$ 以及

$$x_2 = 1.263888889,\quad x_3 = 1.259933493,$$
$$x_4 = 1.259921049,\quad x_5 = 1.259921049.$$

二次收敛是显见的.

插值方法

25.16 这一古老的方法使用前二个逼近值并通过它们之间的一个线性插值构造下一个逼近值，导出试位法(见图 25.5)，

$$c = a - \frac{(a-b)f(a)}{f(a)-f(b)}.$$

解 ☞ 线性函数

$$y = f(a) + \frac{f(a)-f(b)}{a-b}(x-a)$$

图 25.5

明显地在 a 与 b 上取 $y=f(x)$. 在由试位法给出的变量 c 处它为零,这个零可用作我们对 $f(x)=0$ 的根的下一个逼近.在根的邻域中我们是这样有效地将曲线换成了一个线性配置多项式,还要注意的是两个已知的逼近值 a 与 b 是在精确根的两侧的.因此 $f(a)$ 及 $f(b)$ 有相反的符号.当使用试位法时被假设为相反的符号.据此,已找到一个 c, 再次应用试位法时我们在使用这个 c 时或当作新的 a 或是新的 b, 不管那一种选择均维持相反的符号,在图 25.5 中, c 会变成新的 a, 以这种方法可以生成一个逼近序列 $x_0, x_1, x_2, \cdots$, x_0 与 x_1 是原始的 a 及 b.

25.17 应用试位法于 Leonardo 方程.

解 ☞ 选择 $x_0=1$ 及 $x_1=1.5$ 公式产生

$$x_2 = 1.5 - \frac{0.5(2.875)}{9.875} \approx 1.35,$$

$$x_3 = 1.35 - \frac{(-0.15)(-0.3946)}{-3.2696} \approx 1.368,$$

等等,收敛速率可以证明比在题 25.2 中的速率高但不如 Newton 法好.

25.18 一个自然的下一步就是用一个二次的插值多项式而不是一个一次的,假如手边有三个逼近值 x_0, x_1, x_2, 导出一个关于新的逼近值 x_3 的公式,它是这样的一个二次式的零点(根).

解 ☞ 不难证明,通过这三点 (x_0,y_0), (x_1,y_1), (x_2,y_2) 的二次式,其中 $y=f(x)$, 可以写成

$$p(x) = \frac{x_1 - x_0}{x_2 - x_0}(Ah^2 + Bh + C),$$

其中 $h = x - x_2$ 及 A, B, C 是

$$A = \frac{(x_1 - x_0)y_2 + (x_0 - x_2)y_1 + (x_2 - x_1)y_0}{(x_2 - x_1)(x_1 - x_0)^2},$$

$$B = \frac{(x_1 - x_0)(2x_2 - x_1 - x_0)y_2 - (x_2 - x_0)^2 y_1 + (x_2 - x_1)^2 y_0}{(x_2 - x_1)(x_1 - x_0)^2},$$

$$C = \frac{x_2 - x_0}{x_1 - x_0}y_2.$$

就 h 求解 $p(x)=0$ 我们得

$$h = -\frac{2C}{B \pm \sqrt{B^2 - 4AC}}.$$

选择二次公式的这种形式是为了避免在作减法时避免有效数字的损失.这里的符号应选择得使分母的绝对值为较大的,然后

$$x_3 = x_2 + h$$

变成下一个逼近值,这个过程可以重复通过将所有的下标加 1,这个刚刚描述过的方法就是被称作 Muller 方法的.人们发现它收敛到实根也收敛到复根.当然,对后者而言,以算术算法进行,然而即使对于实根,复算法也不失为一种明智的选择,由于虚 P 的踪迹偶尔也会进入计算.

Bernoulli 方法

25.19 证明假如 n 次多项式

$$p(x) = a_0x^n + a_1x^{n-1} + \cdots + a_n$$

有一个单重的最大零点,譬如说 r_1, 然后可以通过计算关于 n 阶差分方程

$$a_0x_k + a_1x_{k-1} + \cdots + a_nx_{k-n} = 0$$

的解序列,并取 $\lim(x_{k+1}/x_k)$ 而得到.

证 ☞ 该差分方程有 $p(x)=0$ 作为它的特征方程,因而它的解可以写成

$$x_k = c_1r_1^k + c_2r_2^k + \cdots + c_nr_n^k,$$

假如我们选择初始值使 $c_1 \neq 0$, 则

$$\frac{x_{k+1}}{x_k} = r_1\frac{1 + (c_2/c_1)(r_2/r_1)^{k+1} + \cdots + (c_n/c_1)(r_n/r_1)^{k+1}}{1 + (c_2/c_1)(r_2/r_1)^k + \cdots + (c_n/c_1)(r_n/r_1)^k}$$

且由于 r_1 为最大的根

$$\lim \frac{r_i}{r_1}=0, \qquad i=2,3,\cdots,n^{*},$$

使 $\lim(x_{k+1}/x_k)=r_1$ 正像所要求的那样. 可以证明使用复变理论初始值 $x_{-n+1}=0=\cdots=x_{-1}=0$, $x_0=1$ 将保证 $c_1\neq 0$.

25.20 应用 Bernoulli 方法于 $x^4-5x^3+9x^2-7x+2=0$.

解 ☞ 与之相关联的差分方程为

$$x_k-5x_{k-1}+9x_{k-2}-7x_{k-3}+2x_{k-4}=0,$$

假如我们取 $x_{-3}=x_{-2}=x_{-1}=0$ 与 $x_0=1$, 则下继的 x_k 在表 25.4 中给出. 同时给定比 x_{k+1}/x_k. 它收敛到 $r=2$ 是慢的. Bernoulli 方法的收敛速率为线性的, 这个方法往往是用来为 Newton 或 Steffensen 方法生成一个好的出发逼近值, 这二种方法均为二次的

表 25.4

k	x_k	x_{k+1}/x_k	k	x_k	x_{k+1}/x_k
1	5	3.2000	9	4,017	2.0164
2	16	2.6250	10	8,100	2.0096
3	42	2.3571	11	16,278	2.0056
4	99	2.2121	12	32,647	2.0032
5	219	2.1279	13	65,399	2.0018
6	466	2.0773	14	130,918	2.0010
7	968	2.0465	15	261,972	2.0006
8	1,981	2.0278	16	524,097	

25.21 为了适应一对共轭复根为主根的情况, 将 Bernoulli 方法加以改动.

解 ☞ 令 r_1 及 r_2 为复共轭根. 则 $|r_i|<|r_1|$ 当 $i=3,\cdots,n$, 由于 r_1, r_2 对是最大的, 用实的出发值, 差分方程的解可以改写为

$$x_k=c_1r_1^k+c_2r_2^k+\cdots+c_nr_n^k,$$

其中 c_1 与 c_2 也是复共轭的. 令 $r_1=re^{-i\phi}=\bar{r}_2, c_1=ae^{i\theta}=\bar{c}_2$, 只有 $r>0, a>0$ 及 $0<\phi<\pi$, 故 r_1 是在上半平面的根. 接着

$$\begin{aligned}x_k&=2ar^k\cos(k\phi+\theta)+c_3r_3^k+\cdots+c_nr_n^k\\&=2ar^k\left[\cos(k\phi+\theta)+\frac{c_3}{2a}\left(\frac{r_3}{r}\right)^k+\cdots+\frac{c_n}{2a}\left(\frac{r_n}{r}\right)^k\right],\end{aligned}$$

除第一项外所有的项有极限为零, 对大的 k 也如此, $x_k\approx 2ar^k\cos(k\phi+\theta)$, 我们现在用这个结果来决定 r 及 ϕ. 首先我们考察到

$$x_{k+1}-2r\cos\phi x_k+r^2x_{k-1}\approx 0$$

正如可以看成从前方程来替代 x_k 并使用关于余弦函数的和与差分的恒等式. 将下标减 1, 我们同样可得到

$$x_k-2r\cos\phi x_{k-1}+r^2x_{k-2}\approx 0.$$

现在同时解这二个式子得

$$r^2\approx\frac{x_k^2-x_{k+1}x_{k-1}}{x_{k-1}^2-x_kx_{k-2}}, \qquad -2r\cos\phi\approx\frac{x_{k+1}x_{k-2}-x_{k-1}x_k}{x_{k-1}^2-x_kx_{k-2}}.$$

决定 r_1 及 r_2 的必要准备现在已具备.

25.22 应用 Bernoulli 方法于 Leonardo 方程.

解 ☞ 相关联的差分方程为 $x_k=-2x_{k-1}-10x_{k-2}+20x_{k-3}$ 而对初始值为 $x_{-2}=x_{-1}=0, x_0=1$

* 译注: 原书此处有误.

的解序列出现在表 25.5 中，对 r^2 及 $-2r\cos\phi$ 的某些逼近值也列在表 25.5 中.

表 25.5

k	x_k	k	x_k	r^2	$-2r\cos\phi$
1	-2	7	2,608	14.6026	3.3642
2	6	8	-32,464	14.6076	3.3696
3	52	9	147,488	14.6135	3.3692
4	-84	10	-22,496	14.6110	3.3686
5	-472	11	-2,079,168	14.6110	3.3688
6	2,824	12	7,333,056		

± 符号的波动是有大复根出现的象征，通过回忆 x_k 的在题 25.21 中给出的形式，即 $x_k \approx 2ar^k\cos(k\phi+\theta)$ 可以看到这一点，当 k 增加时，余弦函数的值在 ±1 之间以依赖于 ϕ 值的大小以有点不规则的方式变化.

由最后一次逼近我们得到

$$r\cos\phi \approx -1.6844, \qquad r\sin\phi = \pm\sqrt{r^2-(r\cos\phi)^2} \approx \pm 3.4313,$$

使得根的主对为 $r_1r_2 \approx -1.6844 \pm 3.4313i$. 由于 Leonardo 方程是三次的，这些根还可以先通过前面所获得的实根将它还原成一个二次方程而得到. 在这种情况下并不真正需要 Bernoulli 方法，所得结果可以通过所有根的总和为 -2，乘积为 20 来检验.

降阶法(Deflation)

25.23 使用单个方程 $x^4-10x^3+35x^2-50x+24=0$ 来说明降阶法思想.

解 这个方程的主根精确地为 4. 应用因子定理我们以除法移掉因子 $x-4$

1	-10	35	-50	24	4
	4	-24	44	-24	
1	-6	11	-6	0	

商为 3 次的 $x^3-6x^2+11x-6$，于是我们说原始的四次多项式下降成为三次的. 三次方程的主根为 3. 将这个因子移去

1	-6	11	-6	3
	3	-9	6	
1	-3	2	0	

我们完成了第二次的下降，降至二次式 x^2-3x+2，它可以将剩下的二个根 2 及 1 解出. 或者二次的还可以下降为线性的 $x-1$. 下降的思想就是指，已经找到了一个根，原始方程可以用低一次的来替代. 理论上，一个寻找方程主根的方法，诸如 Bernoulli 方法，凭借逐次地降阶，它移掉每一个已找到的主根，并假设没有二个根是相同的，这样一个接一个地找到其他所有的根，事实上有误差问题，它限制了这个方法的使用，正如下题所提示的.

25.24 证明若主根的位置不是精确地知道，那么降阶法可能会产生精确度还要差一些的下一个根，并提出一种获得第二个根与第一个根相同精确的方法.

证 假设，为简单起见，所得到的前个方程的主根只准确到二位为 4.005，降阶带来

1	−10	35	50	24	4.005
	4.005	−24.01	44.015	−23.97	
1	−5.995	10.99	−5.985	0.03	

因而三次式是 $x^3-5.995x^2+10.99x-5.985$.这个三次式的主根(准确到两位)是 2.98.至于涉及到原始的四次方程,它的不准确在最后一位上.在这点上,自然的步骤是使用 2.98 作为 Newton 迭代的首次近似,很快地它便会产生一个原始方程的准确到两位的根.这时再作下一次的下降.在实践中发现较小的"根"要求持续地校正,以及甚至对于中等次数的多项式,由下降法所得到的结果对保证 Newton 迭代收敛性所要的那个根而言也可能不是足够好的.当共轭复根 $a\pm bi$ 通过以二次因子 $x^2-2ax+a^2+b$ 去除而移去时同样的注释成立.

商-差分算法

25.25 什么是一个商-差分格式.

解 ☞ 给定一个多项式 $a_0x^n+a_1x^{n-1}+\cdots+a_n$ 及相关联的差分方程

$$a_0x_k+a_1x_{k-1}+\cdots+a_nx_{k-n}=0,$$

考虑取 $x_{-n+1}=\cdots=x_{-1}=0$ 及 $x_0=1$ 的解序列.令 $q_k^1=x_{k+1}/x_k$ 及 $d_k^0=0$.然后定义

$$q_k^{j+1}=\left(\frac{d_{k+1}^j}{d_k^j}\right)q_{k+1}^j,\qquad d_k^j=q_{k+1}^j-q_k^j+d_{k+1}^{j-1},$$

其中 $j=1,2,\cdots,n-1$ 及 $k=0,1,2,\cdots$这些不同的商(q)与差分(d)可以如在表 25.6 中那样列出.通过考察表中菱形部分,定义容易被回忆起来.在一个(q)列为中心的菱形中 SW(南西)对的和与 NE(北东)对的和相等.在一个(d)列为中心的菱形中相应的乘积是相等的.这些是**菱形法则**.

表 25.6

	q_0^1						
0		d_0^1					
	q_1^1		q_0^2				
0		d_1^1		d_0^2			
	q_2^1		q_1^2		q_0^3		
0		d_2^1		d_1^2		d_0^3	
	q_3^1		q_2^2		q_1^3		q_0^4
0		d_3^1		d_2^2		d_1^3	
	q_4^1		q_3^2		q_2^3		q_1^4
0		d_4^1		d_3^2		d_2^3	
	q_5^1	⋮	q_4^2	⋮	q_3^3	⋮	q_2^4
	⋮		⋮		⋮		⋮

25.26 计算与 Fibonacci 序列相关连的多项式的商-差分格式.

解 ☞ 结果列在表 25.7 中.

表 25.7

k	x_k	d_k^0	q_k^1	d_k^1	q_k^2	d_k^2
0	1	0				
			1.0000			
1	1	0		1.0000		

续表

k	x_k	d_k^0	q_k^1	d_k^1	q_k^2	d_k^2
			2.0000		-1.0000	
2	2	0		-0.5000		-0.0001
			1.5000		-0.5001	
3	3	0		0.1667		-0.0001
			1.6667		-0.6669	
4	5	0		-0.0667		0.0005
			1.6000		-0.5997	
5	8	0		0.0250		0.0007
			1.6250		-0.6240	
6	13	0		-0.0096		-0.0082
			1.6154		-0.6226	
7	21	0		0.0037		
			1.6190			
8	34	0				

25.27 什么是与商-差分格式相关联的第一个收敛定理?

解 ☞ 假设给定的多项式没有二个零点有相同的绝对值.于是当 k 趋向无穷时,

$$\lim q_k^j = r_j, \qquad j = 1,2,\cdots,n,$$

其中 $r_1, r_2, \cdots, r_n$ 是按绝对值减少的顺序排列的,当 $j=1$ 时它是 Bernoulli 关于主根的结果.关于 j 的其他值,证明要求复函数理论为而将被省略,这里也曾假设过包含在格式中没有一个分母为零.q 的收敛于根隐含着 d 的收敛于零.这可以在下面看到:从题 25.25 的第一个定义方程,

$$\frac{d_{k+1}^j}{d_k^j} = \frac{q_k^{j+1}}{q_{k+1}^j} \to \frac{r_{j+1}}{r_j} < 1,$$

因此 d_k^j 几何地收敛到零.在目前的这个问题中,收敛的开始在表 25.7 中已经是显然的,除了最后一列将要简短地被讨论外.在这个表中(q)列必须,凭借收敛定理,趋于根 $(1\pm\sqrt{5})/2$,它近似地为 1.61803 及 -0.61803,明显地我们更接近于第一个而不是第二个.

25.28 一个商-差分格式怎样产生一对复共轭根?

解 ☞ 这种根的存在可以用不收敛于零的(d)列来说明.假设 d_k^j 表值的列并不如此,那么我们就构造多项式

$$p_j = x^2 - A_j x + B_j,$$

其中当 k 趋向无穷时有

$$A_j = \lim(q_{k+1}^j + q_k^{j+1}), \qquad B_j = \lim q_k^j q_k^{j+1}.$$

多项式会有根 r_j 及 r_{j+1},它们是复共轭的.从根本上来说,可以得到原始多项式的一个二次因子.这里我们假设 d_k^{j-1} 及 d_k^{j+1} 表值的列正是收敛于零.假如他们不是这样的,那么就有二个以上的根有相同的绝对值.因而需要一个更为复杂的过程.细节,以及刚才所作的收敛性,在国家标准局应用数学序列,49 卷中给出.

25.29 什么是生成一个商-差分格式的行-对-行(row-by-row)方法以及它的好处何在?

解 ☞ 从题 25.25 中首先引入的按列方法对舍入误差十分敏感.这就使表 25.7 的最后一列不像一个(d)列本该那样地收敛于零,而代之以展示一个误差爆炸的典型开始的这一事实得到解释.

下面的按行(row by row)方法对误差则有较小的敏感.虚构表值是用来填补商-差分格式顶端的两行的,其做法如下,从 d_k^0 列出发并以 d_k^n 结束,靠边的两列对所有的 k 均由零组成.这相当于强加在这些边界差分上的适当性态为了控制舍入误差而作的努力.

	$-a_1/a_0$		0		0		0	
0		a_2/a_1		a_3/a_2		a_4/a_3		0

然后应用菱形法则，依次填每个新行，可以证明在题 25.25 中已得到的相同格式将被这个方法所发展，假设在两种方法中都没有误差．在出现误差的情况下逐行方法更为稳定，注意在本法中无需计算 x_k．

25.30 应用行对行方法于 Fibonacci 序列的多项式 x^2-x-1．

解 ☞ 如前题中所提示的那样，填入顶行其他的用菱形法则进行计算．表 25.8 展示这个结果．在最后(q)列中性态的改进是显见的．

表 25.8

k	d	q	d	q	d
		1		0	
1	0		1		0
		2		-1	
2	0		-0.5000		0
		1.5000		-0.5000	
3	0		0.1667		0
		1.6667		-0.6667	
4	0		-0.0667		0
		1.6000		-0.6000	
5	0		0.0250		0
		1.6250		-0.6250	
6	0		-0.0096		0
		1.6154		-0.6154	
7	0		0.0037		0
		1.6191		-0.6191	
8	0				0

25.31 应用商-差分算法来寻找

$$x^4-10x^3+35x^2-50x+24=0$$

的所有的根．

解 ☞ 这个方程的根精确地为 1,2,3,及 4．然而，这个算法并不要求事先知道关于这些根的信息．所以这个方程当作一个简单试验例子．商-差分格式，它由题 25.29 的方法所生成的列在表 25.9 中．明显地，它的收敛是慢的，但是形成了所期望的模式．(d)列看来是走向零的而(q)列按 4,3,2,1 的那种顺序．或许在这时转向 Newton 法是明智的，它非常快地将一个合理的首次逼近，正像我们现在有的，转变为精确结果．商-差分算法正是为了给 Newton 法作准备的这个目的而常被使用．

表 25.9

k	d	q	d	q	d	q	d	q	d
		10		0		0		0	
1	0		-3.5000		-1.4286		-0.4800		0
		6.5000		2.0714		0.9486		0.4800	
2	0		-1.1154		-0.6542		-0.2429		0
		5.3846		2.5326		1.3599		0.7229	
3	0		-0.5246		-0.3513		0.1291		0
		4.8600		2.7059		1.5821		0.8520	
4	0		0.2921		-0.2054		-0.0695		0
		4.5679		2.7926		1.7180		0.9215	
5	0		-0.1786		-0.1264		-0.0373		0
		4.3893		2.8448		1.8071		0.9588	
6	0		-0.1158		-0.0803		-0.0198		0
		4.2735		2.8803		1.8676		0.9786	
7	0		-0.0780		0.0521		-0.0104		0
		4.1955		2.9062		1.9093		0.9890	
8	0		-0.0540		-0.0342		-0.0054		0
		4.1415		2.9260		1.9381		0.9944	

25.32 应用商-差分算法于 Leonardo 方程.

解 ☞ 仍是使用逐行方法,我们生成的格式陈列在表 25.10 中.

表 25.10

k	d	q	d	q	d	q	d
		-2		0		0	
1	0		5		-2		0
		3		-7		2	
2	0		11.6667		0.5714		0
		-8.6667		5.2381		1.4286	
3	0		7.0513		0.1558		0
		-1.6154		-1.6574		1.2728	
4	0		7.2346		-0.1196		0
		5.6192		-9.0116		1.3924	
5	0		-11.6022		0.0185		0
		-5.9830		2.6091		1.3739	
6	0		5.0596		0.0097		0
		-0.9234		-2.4408		1.3642	

收敛是慢的,假设我们在此地打住.第二个(d)列看来简直不像是朝着零的方向前进的,提示 r_1 与 r_2 为复的,正像我们无论如何已经知道的那样.下一个(d)列出现了向零的趋势,提示一个实根,我们知道它靠近 1.369. Newton 法会快速地从我们这儿有的初始估计值 1.3642 产生一个精确根.回到复对上,我们应用题 25.28 的方法,从(q)的头二列我们计算

$$5.6192-9.0116=-3.3924,\qquad (-1.6154)(-9.0116)\approx 14.5573,$$

$$-5.9830+2.6091=-3.3739,\qquad (5.6192)(2.6091)\approx 14.6611,$$

$$-0.9234-2.4408=-3.3642,\qquad (-5.9830)(-2.4408)\approx 14.6033,$$

于是 $A_1\approx-3.3642$ 与 $B_1\approx14.6033$.因此复根近似地由 $x^2+3.3642x+14.6033=0$ 给出,它使这些根为 $r_1,r_2\approx-1.682\pm3.431i$.

Newton 方法使用复算术可以用来改进这些值,然而另一个称之为 Bairstow 的方法在这里将被简短地介绍.在这个题中我们又一次地使用了商差算法来提供所有根的像样的估计值.对一个能完成这一使命的方法不应指望它快速地收敛.然后在适当的地方将它转向一个二次收敛的算法是自然的一步.

Sturm 序列

25.33 定义一个 Sturm 序列.

解 一个函数序列 $f_0(x),f_1(x),\cdots,f_n(x)$ 它在实数轴的一个区间 (a,b) 上满足条件

1. 每个 $f_i(x)$ 是连续的,
2. $f_n(x)$ 的符号不变,
3. 若 $f_i(r)=0$ 则 $f_{i-1}(r)$ 及 $f_{i+1}(r)\neq0$,
4. 若 $f_i(r)=0$ 则 $f_{i-1}(r)$ 及 $f_{i+1}(r)$ 有相反的符号,
5. 若 $f_0(r)=0$ 则当 h 充分小时满足

$$\operatorname{sign}\frac{f_0(r-h)}{f_1(r-h)}=-1,\qquad \operatorname{sign}\frac{f_0(r+h)}{f_1(r+h)}=1,$$

就称为一个 Sturm 序列.

25.34 证明函数 $f_0(x)$ 在区间 (a,b) 上根的个数等于序列 $f_0(a),f_1(a)\cdots f_n(a)$ 的变号数与序列 $f_0(b),f_1(b),\cdots,f_n(b)$ 的变号数之差.

证 当 x 从 a 增加到 b 时在 Sturm 序列中的变号数只能受到一个或更多的有一个零点的函数的影响,由于所有的函数都是连续的,事实上只有 $f_0(x)$ 的一个零点会影响它.因为,假设当 $i\neq0,nf_i(r)=0$ 则由性质 1,3 与 4 下面符号模式对小的 h 是可能的,

	f_{i-1}	f_i	f_{i+1}
$r-h$	+	±	−
r	+	0	−
$r+h$	+	±	−

或

	f_{i-1}	f_i	f_{i+1}
$r-h$	−	±	+
r	−	0	+
$r+h$	−	±	+

在所有的情况下有一个符号改变,因而移动经过这样一个根并不影响符号改变数.由条件 2 函数 $f_n(x)$ 不可能有一个零点,因而我们最终来到 $f_0(x)$.由条件 5,当我们移动经过根 r 时,我们在 f_0 与 f_1 之间,失去一个变号,这就证明了该定理.人们发现这 5 个条件的设计是考虑到了根-计数的性质.

25.35 若 $f_0(x)$ 是一个无重根的 n 次多项式,怎样才能构造一个 Sturm 序列用来数它的根?

解 令 $f_1(x)=f_0'(x)$ 然后应用 Euclidean 算法来构造余下的序列如下:

$$f_0(x)=f_1(x)L_1(x)-f_2(x),$$
$$f_1(x)=f_2(x)L_2(x)-f_3(x),$$
$$\cdots\cdots$$
$$f_{n-2}(x)=f_{n-1}(x)L_{n-1}(x)-f_n(x),$$

其中 $f_i(x)$ 是 $n-i$ 次的而 $L_i(x)$ 为线性的.

这个序列 $f_0(x),f_1(x),\cdots,f_n(x)$ 将是一个 Sturm 序列.为了证明这一点我们首先指出所有的 f_i 均为连续的,由于 f_0 及 f_1 肯定是的,由于 f_n 是一个常数条件 2 成立.相继的两个 $f_i(x)$ 不会同时为零,由于不然的话所有的包含 f_0 与 f_1 的函数序列都会消失而这将隐含着一个重根.这就证明了条件 3,条件 4 是我们定义的方程的一个直接结论,而 5 被满足则由于 $f_1=f_0'$.

假如方法被应用于有重根的多项式,那么所有 $f_i(x)$ 的同时消失会给出它们存在的证据.多项式的降阶法可以移掉重根,因而允许该方法应用于寻找单根.

25.36 应用 Sturm 序列方法来定

$$x^4-2.4x^3+1.03x^2+0.6x-0.32=0$$

所有实根位置.

解 ☞ 把这个多项式记作 $f_0(x)$,我们首先计算导数.由于我们关心的只是不同 $f_i(x)$的符号,通常用一个正的乘数将首项系数规范化是方便的.据此我们以 1/4 乘$f_0(x)$并取

$$f_1(x) = x^3 - 1.8x^2 + 0.515x + 0.15.$$

下一步是以 f_1 除 f_0,人们获得线性的商 $L_1(x) = x - 0.6$,立刻运用不上它,以及一个余项 $-0.565x^2 + 0.759x - 0.23$.在这一点上一个常见的错误是忘了我们要的是负的这个余项.同样将它规范化,我们得到

$$f_2(x) = x^2 - 1.3434x + 0.4071.$$

以 f_2 除 f_1 带来一个线性商 $L_2 = x - 0.4566$ 和一个它的负余项,规范后为

$$f_3(x) = x - 0.6645.$$

最后以 f_3 除 f_2 我们得到余项为 - 0.0440.取其负的并规范化,我们可以选

$$f_4(x) = 1.$$

现在我们有 Sturm 序列并准备好了去找这些根,确认在表 25.11 中陈列的记号是一件简单的事情,它们说明有一根在区间(-1,0)中,一根在(1,2)中而在(0,1)中有两个根.

表 25.11

	f_0	f_1	f_2	f_3	f_4	变号次数
$-\infty$	+	-	+	-	+	4
-1	+	-	+	-	+	4
0	-	+	+	-	+	3
1	-	-	+	+	+	1
2	+	+	+	+	+	0
∞	+	+	+	+	+	0

在这些区间内挑选更多的点,所有的根可以更加精细地被确定.然而,在用商差分算法时在某个时刻后将它移到一个更快收敛的过程,诸如 Newton 法是明智的.一个提供所有实根所在位置的首次估计的方法,像 Sturm 方法所做的,对精细地确定任何一个根是不经济的.在这个例子中根证明为 -0.5,0.5,0.8,及 1.6.

25.37 证明,在得到充分好的初始逼近的条件下.Newton 法将产生前题中方程的所有根.

证 ☞ 下面的图 25.6 展示了多项式的定性性态.明显地任何一个逼近值 $x_0 < -0.5$ 将引出收敛于这个根的序列,由于这样的一个 x_0 已经在曲线的凸的一侧.类似地任何 $x_0 > 1.6$ 将带来向着最大根的收敛.对原来就紧挨着的根就要求精确的出发值,在这个例子中,为了考察怎样将模糊不清的根对分开,可以不要求单根.从图中可见比 0.5 略小的 x_0 将带来到 0.5 的收敛,而一个比 0.8 略大的 x_0 将带来到 0.8 的收敛,由于在这两种情况下我们都是从凸的一侧开始.注意到从 $x_0 = 0.65$ 出发,它是介于两根之间的.这就意味着将跟随一根水平的切线,事实上它导出 $x_1 \approx 5$,在这之后会

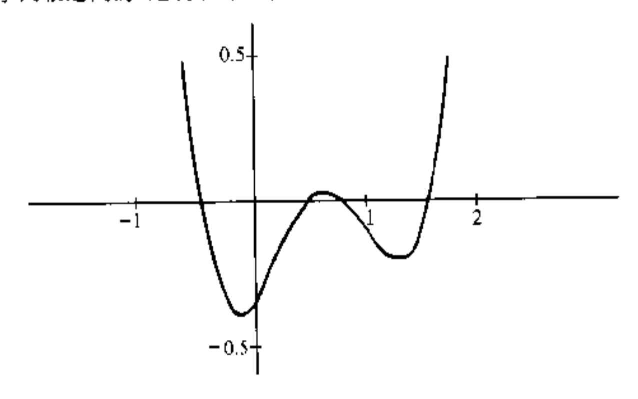


图 25.6

出现收敛到 1.6 处的根，这类情况在 Newton 迭代法中是会出现的.

方程组的 Newton 法

25.38 导出解 $f(x,y)=0, g(x,y)=0$ 的公式

$$x_n = x_{n-1} + h_{n-1},$$
$$y_n = y_{n-1} + k_{n-1},$$

其中 h 和 k 满足

$$f_x(x_{n-1},y_{n-1})h_{n-1} + f_y(x_{n-1},y_{n-1})k_{n-1} = -f(x_{n-1},y_{n-1}),$$
$$g_x(x_{n-1},y_{n-1})h_{n-1} + g_y(x_{n-1},y_{n-1})k_{n-1} = -g(x_{n-1},y_{n-1}),$$

这些公式被称作解两个联立方程的 Newton 方法.

解 ☞ 以 f,g 在 (x_{n-1},y_{n-1}) 邻域的 Taylor 级数的线性部分来逼近它.

$$f(x,y) \approx f(x_{n-1},y_{n-1}) + (x-x_{n-1})f_x(x_{n-1},y_{n-1}) + (y-y_{n-1})f_y(x_{n-1},y_{n-1}),$$
$$g(x,y) \approx g(x_{n-1},y_{n-1}) + (x-x_{n-1})g_x(x_{n-1},y_{n-1}) + (y-y_{n-1})g_y(x_{n-1},y_{n-1}),$$

它假设所包含的导数是存在的.以 (x,y) 表精确解，两个左侧项都为零.定义 $x=x_n$ 及 $y=y_n$ 它们是使得右侧为零的数，我们立得所求的方程.将 Taylor 级数以它的线性部分来替代的思想，就是在题 25.8 中导出解单个方程的 Newton 方法的思想.

25.39 求圆 $x^2+y^2=2$ 与双曲线 $x^2-y^2=1$ 的交点.

解 ☞ 这个特别问题可以用消去法方便地解出.将它们相加得到 $2x^2=3$，于是有 $x\approx\pm1.2247$.相减后得到 $2y^2=1$，于是 $y=\pm0.7071$.知道准确的交点使该问题成为 Newton 方法的一个简单试验例子.取 $x_0=1, y_0=1$.决定 h 与 k 的公式为

$$2x_{n-1}h_{n-1} + 2y_{n-1}k_{n-1} = 2 - x_{n-1}^2 - y_{n-1}^2,$$
$$2x_{n-1}h_{n-1} - 2y_{n-1}k_{n-1} = 1 - x_{n-1}^2 + y_{n-1}^2,$$

而取 $n=1$ 变成 $2h_0+2k_0=0, 2h_0-2k_0=1$.于是

$$h_0 = -k_0 = \frac{1}{4},$$

使 $x_1=x_0+h_0=1.25,\quad y_1=y_0+k_0=0.75$.

下一次迭代带来 $2.5h_1+1.5k_1=-0.125, 2.5h_1-1.5k_1=0$ 使 $h_1=-0.025, k_1=-0.04167$ 及

$$x_2 = x_1 + h_1 = 1.2250, \qquad y_2 = y_1 + k_1 = 0.7083.$$

第三次迭代安排了 $2.45h_2+1.4167k_2=-0.0024, 2.45h_2-1.4167k_2=0.0011$ 使 $h_2=-0.0003$, $k_2=-0.0012$ 及

$$x_3 = x_2 + h_2 = 1.2247, \qquad y_3 = y_2 + k_2 = 0.7071.$$

收敛于准确结果是显见的，可以证明对于充分好的初始逼近值 Newton 方法的收敛是二次的，这个方法的思想可以推广到任何个数的联立方程.

25.40 其他迭代方法也可以对方程组进行推广.例如，若我们的基本方程 $f(x,y)=0$，$g(x,y)=0$可以改写成

$$x = F(x,y), \qquad y = G(x,y),$$

则在对 F 及 G 作适当的假设下，迭代公式

$$x_n = F(x_{n-1},y_{n-1}), \qquad y_n = G(x_{n-1},y_{n-1})$$

对充分精确的初始逼近值收敛.应用这个方法于方程 $x=\sin(x+y), y=\cos(x-y)$.

解 ☞ 这些方程已经是所要求的形式，从平凡的初始逼近值 $x_0=y_0=0$ 出发，我们得到的结果在下面给出.对这样差的逼近值不能把收敛看作是规律性的，往往是对一个已知方程和好的初始逼近值为了去寻找一个收敛的重新安排，人们必须长时间地劳动.

n	0	1	2	3	4	5	6	7
x_n	0	0	0.84	0.984	0.932	0.936	0.935	0.935
y_n	0	1	0.55	0.958	1.000	0.998	0.998	0.998

下降法与优化

25.41　最速下降算法的思想是什么？

解　☞ 各种极小化方法均包含了一个以这样一种方式定义的一个函数 $S(x,y)$，它的极小值精确地出现在 $f(x,y)=0$ 及 $g(x,y)=0$ 的地方，于是解这两个联立方程的问题可以用对 $S(x,y)$ 极小化的问题来替代，例如

$$S(x,y)=[f(x,y)]^2+[g(x,y)]^2$$

肯定在 $f=g=0$ 处达到它的极小值零，这是对 $S(x,y)$ 的通俗的选择．留下来的是如何找到这样一个极小的问题，最速下降法从一个初始逼近值 (x_0,y_0) 开始，在这一点上函数 $S(x,y)$ 在向量

$$-\operatorname{gradient}S(x,y)\big|_{x_0y_0}=[-S_x,-S_y]\big|_{x_0y_0}$$

的方向上下降得最快．为缩写起见记

$$-\operatorname{grad}S_0=[-S_{x0},-S_{y0}].$$

现在以下面的形式得到一个新的近似值

$$x_1=x_0-tS_{x0},\qquad y_1=y_0-tS_{y0},$$

选 t 使 $S(x_1,y_1)$ 为极小值．换言之，我们从 (x_0,y_0) 出发朝着 $-\operatorname{grad}S_0$ 的方向前进，直到 S 开始增长，这样就完成了一步．而另一步从 (x_1,y_1) 开始沿着 $-\operatorname{grad}S_1$ 的方向前进，过程充满希望地一直继续下去直到找到最小点．

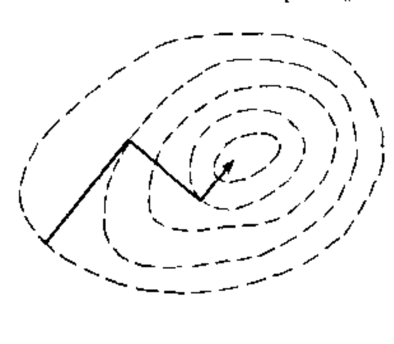


图 25.7

这个过程曾与滑雪者在浓雾中从一座山上回到谷底的过程相比．他不可能看到自己的目标，他开始以最速下降的方向下滑并前进直到他的路径又开始爬坡，然后选一个新的最速下落方向．第二次以同样的方式走．在一个碗-状的为群山所环绕的山谷中，显然这个方法会将它带到离家越来越近的地方．图 25.7 说明这个行为．虚线为等高线或水平线，在该线上 $S(x,y)$ 为常数，在每一点上梯度方向与等高线方向正交．因而我们常以直角离开等高线，沿着这条线前进至 $S(x,y)$ 的最小值意味着来到一个具有更低等高线的切点，事实上它要求无穷多的这类步子到达这个最小值并跟随着一个多少有点不经济的锯齿路程．

25.42　应用一种最速下降的方法来解题 25.40 的方程：

$$x=\sin(x+y),\qquad y=\cos(x-y).$$

解　☞ 这里我们有

$$S=f^2+g^2=[x-\sin(x+y)]^2+[y-\cos(x-y)]^2$$

使

$$\frac{1}{2}S_x=[x-\sin(x+y)][1-\cos(x+y)]$$
$$+[y-\cos(x-y)][\sin(x-y)],$$
$$\frac{1}{2}S_y=[x-\sin(x+y)][-\cos(x+y)]$$
$$+[y-\cos(x-y)][1-\sin(x-y)].$$

假设我们选 $x_0=y_0=0.5$，则 $-\operatorname{grad}S_0=[0.3,0.6]$．由于一个常数倍数可以吸引在参数 t 中，我们可以取

$$x_1=0.5+t,\qquad y_1=0.5+2t.$$

现在要去寻找 $S(0.5+t,.5+2t)$ 的极小．或是直接寻查，或是通过令 $S'(t)$ 为零，我们很快发现极

小值在 $t=0.3$ 附近，使 $x_1=0.8$ 及 $y_1=1.1$. 这个 $S(x_1,y_1)$ 值约为 0.04, 这样我们前进到第二步由于 $-\mathrm{grad}S_1\approx[0.5,-0.25]$. 我们做第一个直角转弯，选

$$x_2=0.8+2t,\qquad y_2=1.1-t,$$

并找 $S(x_2,y_2)$ 的极小值，它证明在 $t=0.07$ 的附近，使 $x_2=0.94$ 与 $y_2=1.03$. 以这种方法持续下去我们得到逐次近似值列表如下，可以指出这个慢收敛是朝着题 25.40 的结果的. 这个方法的慢收敛是典型的，它常被用着对 Newton 法提供好的出发近似值.

x_n	0.5	0.8	0.94	0.928	0.936	0.934
y_n	0.5	1.1	1.03	1.006	1.002	0.998
S_n	0.36	0.04	0.0017	0.00013	0.000025	0.000002

下降法的进行由图 25.8 的途径 A 所提示.

25.43 证明一个最速下降法可以不收敛下所要求的结果.

证 ☞ 使用前题的方程，假设我们选初始逼近 $x_0=y_0=0$, 则 $-\mathrm{grad}S_0=[0,2]$, 于是我们取 $x_1=0$ 及 $y_1=t$. $S(0,t)$ 的极小值证明是在 $t=0.55=y_1$ 处. 以 $S(x_1,y_1)=0.73$. 计算新的梯度，我们得到 $-\mathrm{grad}S_1\approx[-0.2,0]$. 它指点我们离所期望的在 $x=y=1$ 近旁的解朝西走. 接着的步上发现我们顺着在图 25.8 中标着 B 的路径前进. 我们这里的困难对极小化方法是典型的. 在 $x=-0.75$, $y=0.25$ 附近有个次山谷，我们的第一步留给我们的正是到通道的西边或是二个山谷之间的鞍点，因此在 $(0,0.55)$ 处的下降方向是向西的，接着下降进入次山谷. 通常在找到一条成功的路需可观次的试验.

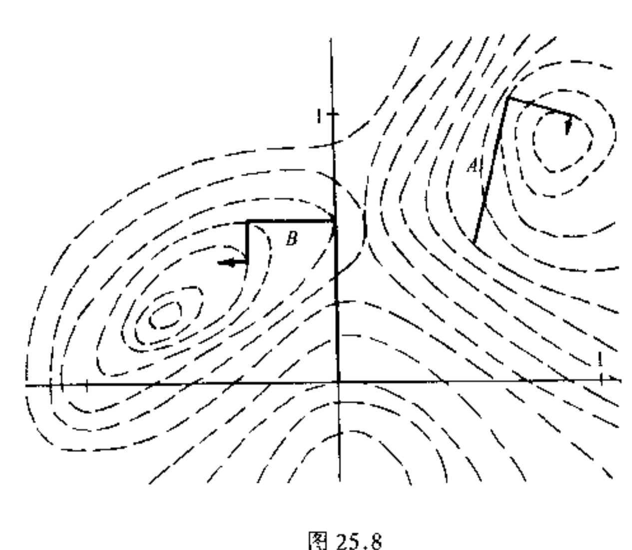


图 25.8

25.44 将下降法的思想推广到解最优化问题或是解非线性方程组.

解 ☞ 两个主要的问题是以什么方向前进以及前进多远. 公式

$$x^{(n)}=x^{(n-1)}+tu_{n-1}$$

对所有选择都开放，$x^{(n-1)}$ 表当前的逼近值，u_{n-1} 是在寻查的下个方向上的一个单位向量. 而 t 就是走多远的度量. 对最速下降法来说，u_{n-1} 就是负梯度向量. 广泛的各种选择都曾被提出过. 理想地，或者人们应跟踪一条曲线，它是等高面的一条正交轨线，在等高面上 f 为常数，其中 f 是要优化的函数. 然而，这就导出了微分方程，使用等步长的最速下降法、等价于应用解微分方程的 Euler 方法. 甚至 Newton 法可以看成一个下降法，取 tu_{n-1} 等于 $-J^{-1}(x^{(n-1)})F(x^{(n-1)})$, 以在引言中用过的记号

二次因子，Bairstow方法

25.45 发展一个递推公式，关于

$$q(x) = b_0x^{n-2} + \cdots + b_{n-2}, \qquad r(x) = b_{n-1}(x-u) + b_n$$

中的系数 b_k．而 $q(x)$ 及 $r(x)$ 由

$$p(x) = a_0x^n + \cdots + a_n = (x^2 - ux - v)q(x) + r(x)$$

所定义．

解 ☞ 将右侧乘出来并比较两侧 x 的幂，我们有

$$b_0 = a_0,$$
$$b_1 = a_1 + ub_0,$$
$$b_k = a_k + ub_{k-1} + vb_{k-2} \quad k = 2, \cdots, n,$$

假如我们人为地令 $b_{-1} = b_{-2} = 0$．最后一个递推式对 $k = 0, 1, \cdots, n$ 成立．这些 b_k 当然依赖于数 u 及 v．

25.46 前题中的递推公式可以怎样地使用于计算 $p(x)$ 在一个复变量 $x = a + bi$ 上的值（假设 a_k 为实的）．

解 ☞ 取 $u = 2a$ 及 $v = -a^2 - b^2$，我们有 $x^2 - ux - v = 0$ 因而 $p(x) = b_{n-1}(x-2a) + b_n$．这一方法的优点是用实算来得到 b_k，所以直到最后一步都不出现复算术．特别，若 $b_{n-1} = b_n = 0$ 则我们有 $p(x) = 0$．复共轭 $a \pm bi$ 就是 $p(x)$ 的零点．

25.47 发展 Bairstow 方法为了使用 Newton 迭代法来解方程组 $b_{n-1}(u, v) = 0, b_n(u, v) = 0$．

解 ☞ 使用 Newton 迭代法，正如在题 25.38 中所描述的那样，我们需要 b_{n-1} 及 b_n 对 u 及 v 的偏导数，首先取对 u 的导数，并令 $c_k = \partial b_{k+1}/\partial u$，我们得到 $c_{-2} = c_{-1} = 0, c_0 = b_0, c_1 = b_1 + uc_0$，接着

$$c_k = b_k + uc_{k-1} + vc_{k-2}$$

最后的这个结果对 $k = 0, 1, \cdots, n-1$ 是实际成立的，因此 c_k 由 b_k 算得，正像 b_k 从 a_k 求得．我们所需的两个结果是

$$\frac{\partial b_{n-1}}{\partial u} = c_{n-2}, \frac{\partial b_n}{\partial u} = c_{n-1}.$$

类似地对 v 求导数并令 $d_k = \partial b_{k+2}/\partial v$ 我们得到 $d_{-2} = d_{-1} = 0$，接着 $d_1 = b_1 + ud_0$，之后

$$d_k = b_k + ud_{k-1} + vd_{k-2},$$

后者对 $k = 0, 1, \cdots, n-2$ 成立．由于 c_k 及 d_k 因此满足相同的递推公式具有相同的初始条件，我们已经证明了，当 $k = 0, 1, \cdots, n-2$，时 $c_k = d_k$ 特别，

$$\frac{\partial b_{n-1}}{\partial v} = c_{n-3}, \qquad \frac{\partial b_n}{\partial v} = c_{n-2}.$$

我们已作好了 Newton 迭代的准备．

假设我们有 $p(x) = 0$ 的近似根 $a \pm bi$，以及与 $p(x)$ 相关联的二次因子 $x^2 - ux - v$．这意味着我们有 $b_{n-1} = b_n = 0$ 的近似根并寻找改进的近似值 $u + h, v + k$．校正量 h 及 k 定义为

$$c_{n-2}h + c_{n-3}k = -b_{n-1},$$
$$c_{n-1}h + c_{n-2}k = -b_n.$$

这些是 Newton 迭代的中心方程，解出 h 与 k，

$$h = \frac{b_nc_{n-3} - b_{n-1}c_{n-2}}{c_{n-2}^2 - c_{n-1}c_{n-3}}, \qquad k = \frac{b_{n-1}c_{n-1} - b_nc_{n-2}}{c_{n-2}^2 - c_{n-1}c_{n-3}}.$$

25.48 应用 Bairstow 方法来决定 Leonardo 方程的准确到 9 位的复根．

解 ☞ 我们已经用商-差分算法得到极好的初始逼近值（参看题 25.32）：$u_0 \approx -3.3642$，$v_0 \approx -14.6033$．我们的递推公式现在产生下面的 b_k 及 c_k：

k	0	1	2	3
a_k	1	2	10	−20
b_k	1	−1.3642	−0.01386	−0.03155
c_k	1	−4.7284	1.2901	

接着题 25.47 产生 $h=-0.004608, k=-0.007930$ 使得

$$u_1=u_0+h=-3.368808,\qquad v_1=v_0+k=-14.611230.$$

重复这个过程,我们接着寻找新的 b_k 及 c_k:

k	0	1	2	3
a_k	1	2	10	−20
b_k	1	−1.368808	0.000021341	−0.000103380
c_k	1	−4.737616	1.348910341	

这些带来

$$h=-0.000000108,\qquad k=-0.000021852,$$

$$u_2=-3.368808108,\quad v_2=-14.611251852.$$

再一次重复这个循环得到 $b_2=b_3=h=k=0$ 到 9 位.现在所要求的根为

$$x_1,x_2=\frac{1}{2}u\pm i\sqrt{-v-\frac{1}{4}u^2}$$

$$=-1.684404054\pm 3.431331350i.$$

这些可以进一步地通过计算所有三个根的和与乘积,并与 Leonardo 方程的系数 2 及 20 进行比较得到检验.

补　充　题

25.49　应用题 25.1 的方法于方程 $x=e^{-x}$ 求一个接近 $x=0.5$ 的根.证明从 $x_0=0.5$ 出发,逼近值 x_{10} 及 x_{11} 在 0.567 处一致到三位.

25.50　对前题计算出来的前面的逼近值进行 Aitken 加速.问何时它产生三位精确值?

25.51　将方程 $x^3=x^2+x+1$ 改写成 $x=1+1/x+1/x^2$,然后使用如题 25.1 中那样的一种迭代法来求一个正根.

25.52　应用 Newton 法于题 25.49 的方程.达到三位精确需要多少次迭代?达到六位精确又需要多少次迭代呢?

25.53　应用 Newton 方法于题 25.51 的方程.

25.54　求 3 的平方根到 6 位精确.

25.55　求 3 的 5 次根到 6 位精确.

25.56　证明 Newton 法应用于 $f(x)=\frac{1}{x}-Q=0$ 导出关于产生倒数而无需作除法的迭代公式 $x_n=x_{n-1}(2-Qx_{n-1})$.应用这个迭代取 $Q=e\approx 2.7182818$,从 $x_0=0.3$ 开始及再以 $x_0=1$ 开始.这些初始逼近中的一个对产生一个收敛序列的准确结果是不够接近的.

25.57　对题 25.49 的方程应用试位法,从逼近值 0 与 1 开始.

25.58　应用题 25.18 的方法(二次插值)于题 25.49 的方程.

25.59　应用二次插值法于 Leonardo 方程.

25.60　使用 Bernoulli 方法求 Fibonacci 方程 $x^2-x-1=0$ 的主(实)根.

25.61　应用 Bernoulli 方法于题 25.31 的方程.

25.62　应用 Bernoulli 方法求 $4x^4+4x^3+3x^2-x-1=0$ 共轭复根的一个主对.

25.63　使用商-差分方法来求题 25.36 中方程的所有根.

25.64　使用商-差分方法来确定题 25.62 中方程的所有根的位置.

25.65 使用一个 Sturm 序列证明 $36x^6+36x^5+23x^4-13x^3-12x^2+x+1=0$ 只有 4 个实根并确定这 4 个根的位置，然后应用 Newton 法去精化它.

25.66 使用一个 Sturm 序列证明 $288x^5-720x^4+694x^3-321x^2+71x-6=0$ 有 5 个紧靠的实根.应用 Newton 法决定这些根到 6 位.

25.67 使用迭代法求

$$x=0.7\sin x+0.2\cos y \qquad y=0.7\cos x-0.2\sin y$$

的解靠近(0.5,0.5)

25.68 应用 Newton 法于前题的方程组.

25.69 应用 Newton 法于方程组 $x=x^2+y^2, y=x^2-y^2$ 求一个接近(0.8,0.4)的一个解.

25.70 应用最速下降法于前题的方程组.

25.71 应用最速下降法于题 25.67 的方程组.

25.72 已知 1 为 $x^3-2x^2-5x+6=0$ 的精确根，通过降阶法使之为一个二次方程，求它的另外两个根.

25.73 求 $x^4+2x^3+7x^2-11=0$ 的所有的根准确到 6 位.使用降阶法以 Newton 及 Bairstow 迭代作支撑.

25.74 利用 Bairstow 方法于 $x^4-3x^3+20x^2+44x+54=0$ 求一个靠近 x^2+2x+2 的二次因子.

25.75 求 $x^4-2.0379x^3-15.4245x^2+15.6696x+35.4936=0$ 的最大根.

25.76 求 $2x^4+16x^3+x^2-74x+56=0$ 靠近 $x=1$ 的二个根.

25.77 求 $x^3=x+4$ 的任何实根.

25.78 求 $x^{1.8632}=5.2171x-2.1167$ 的一个小正根.

25.79 求 $x=2\sin x$ 的在 $x=2$ 附近的根.

25.80 求 $x^4-3x^3+20x^2+44x+54=0$ 的一个具有负的实部的复根对.

25.81 求方程组

$$x=\sin x\cosh y, \qquad y=\cos x\sinh y$$

的在 $x=7, y=3$ 附近的解.

25.82 在 $x=2, y=3$ 附近解 $x^4+y^4-67=0, x^3-3xy^2+35=0$.

25.83 对正 x 寻找 $y=(\tan x)/x^2$ 的最小值.

25.84 曲线 $y=e^{-x}\log x$ 在何处有一个弯点?

25.85 求 $1-x+\frac{x^2}{(2!)^2}-\frac{x^3}{(3!)^2}+\frac{x^4}{(4!)^2}-\cdots=0$ 的最小正根.

25.86 已知 $\sin(xy)=y-x$.求 $y(x)$ 在 $x=1$ 附近的极大值.

25.87 求 $x^4-x=10$ 在 $x=2$ 附近的一个根到 12 位.

25.88 求 $e^{-x}=\sin x$ 的最小实根.

25.89 将 4 次多项式 $x^4+5x^3+3x^2-5x-9$ 拆成二个二次因子.

25.90 求 $x=\frac{1}{2}+\sin x$ 在 1.5 附近的一个根.

25.91 求 $2x^3-13x^2-22x+3=0$ 的所有根.

25.92 求 $x^6=x^4+x^3+1$ 在 1.5 附近的一个根.

25.93 求 $x^4-5x^3-12x^2+76x-79=0$ 在 $x=2$ 附近的二个根.

25.94 证明通过平移变换 $x=y-a/3$ 将二次项从一个一般的三次方程

$$x^3+ax^2+bx+c=0$$

中移去.同时参看下题.

25.95 在 1545 年 Cardano 发表了这个解 3 次方程 $x^3+bx+c=0$ 的公式(注意缺二次项)

$$x=\left[-\frac{c}{2}+\sqrt{\left(\frac{c}{2}\right)^2+\left(\frac{b}{3}\right)^3}\right]^{1/3}-\left[\frac{c}{2}+\sqrt{\left(\frac{c}{2}\right)^2+\left(\frac{b}{3}\right)^3}\right]^{1/3}.$$

应用它求 $x^3+3x-4=0$ 的至少一个实根 $x=1$，它也能求 $x^3-15x-4=0$ 的实根 $x=4$ 吗?

第二十六章 线性方程组

线性方程组的解

这完全能有理由作为数值分析的首要问题.许多应用数学问题归结为一组线性方程,或一个线性方程组

$$Ax = b$$

具有已知的矩阵 A 与向量 b 以及待定的向量 x.为解它大量的算法已经被推出,其中的若干种算法将在这里提出来.可用算法的多样性表明这问题表面上的初等特征是迷惑人的,存在着数不清的陷阱.

Gauss 消去法是最古老的算法之一并且依然是最流行的算法之一.它包含以方程某种方式的组合来替代原方程使得得到一个三角形方程组.

$$\begin{aligned} u_{11}x_1 + u_{12}x_2 + \cdots + u_{1n}x_n &= c_1, \\ u_{22}x_2 + \cdots + u_{2n}x_n &= c_2, \\ &\vdots \\ u_{nn}x_n &= c_n. \end{aligned}$$

在这之后,x 的分量容易被求得,一个分量跟在另一个之后,通过称作回代的过程.最后一个方程确定 x_n,把它代入次后一个方程而得 x_{n-1}等等.

Gauss 算法还提供对矩阵 A 的因子分解,其形式为 $A = LU$,其中 U 为一个正三角阵如上面所示,而 L 为一个对角元均为 1 的下三角阵.这个算法可以用来证明代数学的基本定理,它处理方程组是否有一个解存在的问题.当相应的齐次方程组 $Ax = 0$ 只有解 $x = 0$ 时定理保证了 $Ax = b$ 确切地有惟一的解.方程组以及系数矩阵 A 这时被称作非奇异的.当 $Ax = 0$ 有其他的不同于 $x = 0$ 的解,则称方程组与矩阵 A 为奇异的.在这种情况下 $Ax = b$ 不是根本没有解,就是有无穷多个解.奇异方程组出现在特征值问题中.假若本章的方法漫不经心地用在一个奇异方程组中,就有令人惊讶的可能性出现,即不可避免的舍入误差将把它变成一个"几乎等同"的非奇异方程组,这时可能产生一个计算"解",而其实根本是不存在的.

因子分解法将 A 转化为LU 或LDU 形式的乘积,其中 L 是一个其主对角线上方的元素为零的矩阵,U 则在主对角线下方的为零,而 D 只有对角线上的元素不为零.矩阵 L 称为下三角阵而 U 则为上三角阵.若 L 或 U 的所有对角元素均为 1,它就被称为单位三角阵. Doolittle, Crout, Cholesky 等方法以及像已经提到的 Gauss 法都产生因子分解.当 A 已以这种方式分解成因子时,解就容易得到了.由于

$$Ax = LUx = L(Ux) = Ly = b,$$

所以我们先就 $Ly = b$ 解出 y,然后就 $Ux = y$ 解出 x.这二个三角方程组中的第一个应对向前回代,而第二个则应对向后回代.

迭代法生成对解向量 x 的逐次逼近序列.这种类型的典型代表是 Gauss-Seidel 方法,它赋予方程组 $Ax = b$ 以如下新形式

$$\begin{aligned} x_1 &= \cdots, \\ x_2 &= \cdots, \\ &\vdots \\ x_n &= \cdots. \end{aligned}$$

让 x_i 从第 i 个方程中解出.一个对所有 x_i 的初始逼近允许每个分量依次得到校正,当一轮循环完成后就开始另一轮.一系列收敛性定理已得到证明.这个方法常用于稀疏矩阵,在那里许

多元素均为零.

使用由

$$r = b - Ax^{(1)}$$

定义的残量向量 r 作迭代改进是一种常用的算法. 令 e 为误差

$$e = x - x^{(1)},$$

并观察

$$Ae = Ax - Ax^{(1)} = b - (b - r) = r,$$

解 $Ae = r$ 产生一个 e 的逼近值, 称它为 $e^{(1)}$, 由此

$$x^{(2)} = x^{(1)} + e^{(1)}$$

成为一个对真解 x 的新的逼近值. 只要看来是有成效的这个过程就可以持续下去.

有广泛的一类更为精细的迭代方法.

在一个计算解 $x^{(c)}$ 中误差的出现其原因是多方面的. 输入数据可能是不完美的, 也就是说, A 及 b 的元素可能会有误差. 在求解算法的执行过程中几乎肯定地会造成舍入误差, 在一个大型问题中它们也许是数以百万计的. 当一个收敛的迭代过程终止时, 手边的逼近值未必就是真解. 这类来源的最后误差可以作出估计, 虽然常常是十分保守的, 然而它们是重要的. 向后误差分析在考察内部舍入问题时是一个有用的工具.

系数矩阵 A 的特性严重地影响误差性态. 接近奇异的方程组对 A 与 b 中的即使是小误差以及对内部的舍入都极其敏感. A 的条件可以使用矩阵范数的概念数值地描述, 大的条件数意味着一个接近奇异的矩阵与相对差的误差控制. 这类矩阵也称为病态的(ill-conditioned, 或称坏条件的). 有时坏条件将被算法的怪异举止而揭示出来. 不幸的是, 这并不总是如此.

矩阵求逆

当然, 知道矩阵的逆, 将会允许 $Ax = b$ 作为一个额外的结果被解出, 因为

$$x = A^{-1}b,$$

然而对线性方程组的求解而言它通常是一个不经济的过程. A^{-1}的元素的完全的知识只是在少数几类应用中被要求, 首推统计分析. 刚才讨论过的解 $Ax = b$ 的方法可以被采用来求逆. 消元, 分解, 迭代以及一个变换法将在问题中加以说明.

特征值问题

特征值问题要求决定数 λ 使线性方程组 $Ax = \lambda x$ 有除了 $x = 0$ 外的其他的解. 这些数称作特征值. 相应的解, 或特征向量, 同样是令人感兴趣的. 列出三种一般性的处理方法.

1. 一个矩阵 A 的特征多项式有 A 的特征值作为它的零点. 为寻求这个多项式类似于 Gauss 消去法的一种直接方法将被包括在内. 可以用 25 章的方法来求它的零点. 以得到的特征值代入 $Ax = \lambda x$ 便产生一个奇异方程组. x 的某些分量值可以被指定, 而简化的方程组可以用我们关于线性方程组的方法来求解.
2. 幂方法生成向量, 以一个多少有点任意的初始向量 V,

$$x^{(p)} = A^{(p)}V,$$

并产生主特征值与它的特征向量. 对于大的 p 值它证明 $x^{(p)}$接近于一个特征向量, 它对应于

$$\lambda = \frac{x^{(p)T}Ax^{(p)}}{x^{(p)T}x^{(p)}},$$

一个称作 Rayleigh 商的公式. 通过修正引出绝对地最小的与某个次大的(next-dominant)特征向量

一个有意义的变招是使用特征值移位的思想来加速幂法的收敛. 逆幂法及逆迭代法都是这种思想的发展.

3. 还原到正则形式(简化后的形式诸如对角矩阵,三对角阵,三角阵,Hessenberg 阵)能用多种方式来进行.当采用相似变换来完成时特征值不变.Jacobi 方法针对一个实对称矩阵进行基于子矩阵

$$\begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

的旋转,并导出一个几乎对角矩阵.Givens 方法使用类似的旋转并以有限的几步达到一个三对角阵.QR 方法,在某种环境下产生一个三角矩阵.所有这些方法的根本思想是正则形矩阵的特征值更容易被求得.

复方程组

假如一个计算机具有复算术功能可用,则许多用来解实方程组的方法可以取来用于复的.假如不然,复方程组可以改变成等价的、但是更大的实方程组.因此比较

$$(A+iB)(x+iy)=a+ib$$

的实部和虚部,导出

$$\begin{bmatrix} A & -B \\ B & A \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}=\begin{bmatrix} a \\ b \end{bmatrix},$$

对它应用我们的实算法.求逆问题

$$(A+iB)(C+iD)=I$$

可以类似的处理.特征值也可用这种方式来处理.

题　解

Gauss 消去法

26.1　用 Gauss 消去法解

$$x_1+\frac{1}{2}x_2+\frac{1}{3}x_3=1,$$
$$\frac{1}{2}x_1+\frac{1}{3}x_2+\frac{1}{4}x_3=0,$$
$$\frac{1}{3}x_1+\frac{1}{4}x_2+\frac{1}{5}x_3=0.$$

解　☞ 我们从在第 1 列中找绝对值最大的系数开始.这里它在顶部.若不然,需要作一个行交换来安排它.这个最大的元素称为第一主元.现在定义

$$l_{21}=\frac{a_{21}}{a_{11}}=\frac{1}{2},\qquad l_{31}=\frac{a_{31}}{a_{11}}=\frac{1}{3}.$$

并将第 1 列的下面二个元素以一种熟悉的方法化为零,即从第 i 个方程减去第一个方程乘以 l_{i1}, $i=2,3$.其结果如下:

$$x_1+\frac{1}{2}x_2+\frac{1}{3}x_3=1,$$
$$\frac{1}{12}x_2+\frac{1}{12}x_3=-\frac{1}{2},$$
$$\frac{1}{12}x_2+\frac{4}{45}x_3=-\frac{1}{3}.$$

这是第一次改变过的方程组.同样的方法现在应用于由下二个方程组成的较小的方程组.又一次地绝对值最大的系数已经在首列的顶部,所以不需要进行行交换.我们找到

$$l_{32}=\frac{a_{32}^{(1)}}{a_{22}^{(1)}}=1,$$

于是从第三个方程减去第二个方程乘以 l_{32}.[上标(1)指第一次变换过的方程组.]于是我们有

$$x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3 = 1,$$
$$\frac{1}{12}x_2 + \frac{1}{12}x_3 = -\frac{1}{2},$$
$$\frac{1}{180}x_3 = \frac{1}{6}.$$

并显然是三角方程组.接下来通过回代来进行求解,它以逆序由底向上求出各分量 x_i:

$$x_3 = 30, x_2 = -36, \qquad x_1 = 9.$$

26.2 为什么主元重要?

解 ☞ 考虑极端的例子:

$$10^{-5}x_1 + x_2 = 1,$$
$$x_1 + x_2 = 2.$$

这个非常小的系数使得解显然应该十分地接近 $x_1 = x_2 = 1$.假设我们不取主元并且只能用4位小数进行计算,精确相减会产生方程

$$(1 - 10^5)x_2 = 2 - 10^5.$$

但是以小数位数的限制我们必须整理成

$$10^5 x_2 = 10^5,$$

它还是以 $x_2 = 1$ 提供给我们,然而,继续回代我们就面对

$$10^{-5}x_1 + 1 = 1$$

造成 $x_1 = 0$ 而不是期待的1

但是现在交换这二个方程,将第1列的最大系数带至主元的位置:

$$x_1 + x_2 = 2,$$
$$10^{-5}x_1 + x_2 = 1.$$

精确的减法现在会带来

$$(1 - 10^{-5})x_2 = 1 - 2(10^{-5}),$$

对于它相同的限制会舍入到 $x_2 = 1$.这时回代成为

$$x_1 + 1 = 2,$$

因而 $x_1 = 1$.选主元造成无意义与完美结果之间的差别.从许多不是那么极端的方程组的经验已经表明选主元是消去法算法的一个重要部分.所描述的技巧称为部分选主元,因为对最大系数的寻查只限于在当前的那列中.进入其他列广泛的寻查将导致列的交换,它的价值尚待讨论.

手边的例子可以用来作进一步的说明.第一个方程乘以 10^5 可以得到

$$x_1 + 10^5 x_2 = 10^5,$$
$$x_1 + x_2 = 2,$$

它使选主元变得没有必要.通常的减法当精确地完成后便得到

$$(1 - 10^5)x_2 = 2 - 10^5.$$

但在舍入后又变成

$$-10^5 x_2 = -10^5,$$

所以 $x_2 = 1$.但是这么一来 $x_1 = 10^5 - 10^5 = 0$,因而我们得到前面的"解".这要点是,当在某处出现非常大的系数时甚至连选主元都可能无济于事.一个走出这个困境的途径可能是列交换,然而替代的办法是将每个方程规格化,使绝对值最大的系数在每个方程中大约相同.做这件事的一个流行的办法是将每个方程均以它的尺寸绝对值为最大的系数来除.每个方程的"范数"将是1.当然在我们的例子中我们会回到原来的方程组.它告诉我们的是,规格化与部分选主元相结合是产生好结果的好机会.

26.3 对 $n \times n$ 线性方程组综述 Gauss 算法.

解 ☞ 假设在题26.1中描述过的类型已进行了 k 步,将方程组带到形式

$$u_{11}x_1 + u_{12}x_2 + \cdots + u_{1k}x_k + u_{1,k+1}x_{k+1} + \cdots + u_{1n}x_n = b'_1,$$
$$u_{22}x_2 + \cdots + u_{2k}x_k + u_{2,k+1}x_{k+1} + \cdots + u_{2n}x_n = b'_2,$$
$$\vdots$$
$$u_{kk}x_k + u_{k,k+1}x_{k+1} + \cdots + u_{kn}x_n = b'_k,$$

$$a_{k+1,k+1}^{(k)}x_{k+1}+\cdots+a_{k+1,n}^{(k)}x_n=b_{k+1}^{(k)},$$
$$\vdots$$
$$a_{n,k+1}^{(k)}x_{k+1}+\cdots+a_{nn}^{(k)}x_n=b_n^{(k)}.$$

顶端的 k 个方程是它们的最终形式，以 $u_{11},\cdots,u_{kk}$ 表头 k 个主元. 在剩下的 $n-k$ 个方程中系数加以这改变后的方程的上标(k). 我们下一步是在底下的 $n-k$ 个方程中的 x_{k+1} 之系数间找第($k+1$)个主元. 它将是绝对值最大的并且它的方程将与第 $k+1$ 个方程交换. 以这个在应放的位置上称其为 $u_{k+1,k+1}$ 的新主元，得到一组新的乘数为

$$l_{i,k+1}=\frac{a_{i,k+1}^{(k)}}{u_{k+1,k+1}},\qquad i=k+2,\cdots,n,$$

通过方程相减，在新的主元下方就被化成零. 系数按以下规则改变：

$$a_{ij}^{(k+1)}=a_{ij}^{(k)}-l_{i,k+1}a_{k+1,j}^{(k)},\qquad k=0,\cdots,n-2,$$
$$b_i^{(k+1)}=b_i^{(k)}-l_{i,k+1}b_k^{(k)},\qquad j=k+2,\cdots,n,\quad i=k+2,\cdots,n,$$

以 $k=0$ 指原方程. 算法的回代部分由

$$x_i=\frac{1}{u_{ii}}\Big(b_i'-\sum_{j=i+1}^{n}u_{ij}x_j\Big),\qquad i=n,\cdots,1$$

表示.

26.4　什么是 Gauss-Jordan 变种?

解　☞ 这里通过进一步相减在每个主元的上方及下方均生成零. 因此最后的矩阵是对角的而不是三角的，所以回代就被取消. 这个思想是吸引人的，然而它包含了比原来的算法更多的计算量因而很少有使用它.

26.5　估计对 $n\times n$ 方程组执行 Gauss 算法所需的计算量.

解　☞ 考虑系数矩阵 A 约化成一个三角形式. 这是工作量占最大份额的地方. 在第一步上，要得到$(n-1)^2$ 个修改后的系数. 我们进一步将我们的注意力限制到这类系数的数量上. 在相继的步中，这个数目在减少但大的总数将是

$$(n-1)^2+(n-2)^2+\cdots+1$$

个系数. 据代数学的一个知名结果它等于$(2n^3-3n^2+n)/6$，从它里面抽出主项 $n^3/3$ 作为一个计算量大小的简单度量. 若 $n=100$，这个数就往 6 位数前进.

26.6　应用 Gauss 消去法于下面方程组，假设计算机只能用二位浮点数进行计算.

$$x_1+67x_2+0.33x_3=2,$$
$$0.45x_1+x_2+0.55x_3=2,$$
$$0.67x_1+0.33x_2+x_3=2.$$

解　☞ 以 $l_{21}=0.45$ 及 $l_{31}=0.67$，下面左侧的数组概括了过程的第一阶段，然后以 $l_{32}=-0.17$ 在右侧的数组表示最后的三角化.

1	0.67	0.33	2.0	1	0.67	0.33	2.0
0	0.70	0.40	1.1	0	0.70	0.40	1.1
0	−0.12	0.78	0.7	0	0	0.85	0.89

回代现在从

$$x_3=\frac{0.89}{0.85}=1.047$$

开始，这里假定我们用双倍精度的累加器，但无论如何要舍入到 1.0. 于是

$$x_2=\left(\frac{1}{0.7}\right)(1.1-0.4)=1.0,$$
$$x_1=2-0.67-0.33=1.0.$$

不顾计算机的严格限制，还是得到精确解(1,1,1). 这是因为我们有一个非常如人意的矩阵.（同时参看题 26.20.）

26.7 Gauss 消去法与系数矩阵的因子之间有什么联系？

解 ☞ 使用题 26.1 的结果，形成矩阵 L 与 U 如下：

$$L=\begin{bmatrix}1&0&0\\l_{21}&1&0\\l_{31}&l_{32}&1\end{bmatrix}=\begin{bmatrix}1&0&0\\\frac{1}{2}&1&0\\\frac{1}{3}&1&1\end{bmatrix},$$

$$U=\begin{bmatrix}u_{11}&u_{12}&u_{13}\\0&u_{22}&u_{23}\\0&0&u_{33}\end{bmatrix}=\begin{bmatrix}1&\frac{1}{2}&\frac{1}{3}\\0&\frac{1}{12}&\frac{1}{12}\\0&0&\frac{1}{180}\end{bmatrix},$$

于是

$$LU=\begin{bmatrix}1&\frac{1}{2}&\frac{1}{3}\\\frac{1}{2}&\frac{1}{3}&\frac{1}{4}\\\frac{1}{3}&\frac{1}{4}&\frac{1}{5}\end{bmatrix}=A.$$

关于这个因子分解的一般证明参看下一题.

26.8 证明若 L 为一个下三角阵具有元素 l_{ij} 以及 $l_{ii}=1$，并且若 U 是一个上三角矩阵具有元素 u_{ij}，则 $LU=A$.

证 ☞ 证明包含了三角矩阵的某些容易的练习.简要地回到开放的例子，定义

$$S_1=\begin{bmatrix}1&0&0\\-\frac{1}{2}&1&0\\-\frac{1}{3}&0&1\end{bmatrix},\qquad S_2=\begin{bmatrix}1&0&0\\0&1&0\\0&-1&1\end{bmatrix}.$$

当它应用到方程的左侧时，考察到乘积 S_1A 实现 Gauss 算法的第一步，而 S_2S_1A 接着实现第二步.这意味着

$$S_2S_1A=U,\qquad A=S_1^{-1}S_2^{-1}U=LU,$$

其中 $L=S_1^{-1}S_2^{-1}$.同时注意到

$$S_1^{1}=\begin{bmatrix}1&0&0\\\frac{1}{2}&1&0\\\frac{1}{3}&0&1\end{bmatrix},\qquad S_2^{-1}=\begin{bmatrix}1&0&0\\0&1&0\\0&1&1\end{bmatrix},$$

因而求逆通过改变 l_{ij} 元的符号而实现.

对于一般的问题先假设无需交换.定义矩阵

$$L_i=\begin{bmatrix}1&&&&\\&\ddots&&&\\&&1&&\\&&-l_{i+1,i}&&\\&&\vdots&\ddots&\\&&-l_{n,i}&&1\end{bmatrix},\quad i=1,\cdots,n-1,$$

所有其他元素为零.正如在例子中那样，这些 L_i 中的每一个实现消元过程的一步，使

$$L_{n-1}L_{n-2}\cdots L_1A=U,$$

这意味着

$$A=L_1^{-1}\cdots L_{n-1}^{-1}U=LU,$$

由于具有对角元为 1 的下三角阵的乘积本身为同类型的，故我们得到因子分解.此外，由于每一个逆

是通过改变 l_{ij} 元的符号来实现的，所以这些都是容易做到的并可以通过相乘来再现

$$L=\begin{bmatrix}1 & 0 & \cdots & & 0\\ l_{21} & 1 & \cdots & & 0\\ \cdots & & & & \\ l_{n1} & l_{n2} & \cdots & l_{n,n-1} & 1\end{bmatrix}.$$

现在假设要作某些行交换．引进交换矩阵

$$I_{ij}=\begin{bmatrix}1 & & & & & \\ & \ddots & & & & \\ & & 0 & 1 & & \\ & & 1 & 0 & & \\ & & & & \ddots & \\ & & & & & 1\end{bmatrix}\begin{matrix}\\ \\ i\text{ 行}\\ j\text{ 行}\\ \\ \\ \end{matrix},$$

$$i\text{ 列}\quad j\text{ 列}$$

乘积 $I_{ij}A$ 会将 A 的 i 和 j 行交换，而 AI_{ij} 会将相应列交换，消去法现在使用了一连串的 I_{ij} 交换和 L_i 运算，导出了表达式

$$L_{n-1}I_{n-1,r_{n-1}}L_{n-2}I_{n-2,r_{n-2}}\cdots L_1L_{1,r_1}A=U,$$

其中 r_i 为含有被选主元的行．它可以重新整理成

$$(L_{n-1}L_{n-2}\cdots L_1)(I_{n-1,r_{n-1}}\cdots I_{1,r_1})A=U$$

或

$$L^{-1}PA=U,\qquad PA=LU.$$

以 P 表包括 $n-1$ 个交换的置换矩阵．假设 A 非奇异，这意味着存在着一个行置换使 PA 有 LU 因子分解．这个分解的惟一性由题 26.14 看将是显然的．

26.9　假设 LU 因子分解已经完成求解方程组 $Ax=b$．

解　☞ 由于 L,U 及 P 均在握，我们有

$$Ax=LUx=PAx=Pb,$$

并令 $y=Ux$，首先从 $Ly=Pb$ 来解 y．它可用向前回代方便地完成．然后以向后回代解 $Ux=y$．更加明确，取 p_i 表 Pb 的一个元素，方程组 $Ly=Pb$ 为

$$\begin{aligned}&l_{11}y_1 &&=p_1,\\ &l_{21}y_1+l_{22}y_2 &&=p_2,\\ & &&\ \vdots\\ &l_{n1}y_1+l_{n2}y_2+\cdots+l_{nn}y_n &&=p_n,\end{aligned}$$

其中所有 $l_{ii}=1$．通过向前回代解显然为 $y_1=p_1,y_2=p_2-l_{21}y_1$，或更为一般地有，

$$y_r=p_r-l_{r1}y_1-\cdots-l_{r-1,r-1}y_{r-1},$$

对 $r=1,\cdots,n$．然后以题 26.3 的公式完成向后回代，变化的只是将 b' 换成 y：

$$x_i=\left(\frac{1}{u_{ii}}\right)(y_i-u_{i,i+1}x_{i+1}-\cdots-u_{in}x_n)$$

以 $i=n,\cdots,1$．如果方程组必须有不止一个 b 要解时，分解因子与向前-向后回代的组合特别有用．

26.10　什么是一个紧致的算法？

解　☞ 当 Gauss 消去法用手算来完成，A 的许多元素被重复多次．在计算机中这会等价于造成存储空间的无节制使用．对于用大型方程组可取的是存储空间及计算机时间两方面都要经济．为此，设计了紧凑算法．例如，消元过程，矩阵 A 的下三角被零所置换．这些存储的位置可以更好地用来记录逐次的 $j<i$ 的 l_{ij} 值．在执行结束时 A 的上三角将被 U 所取代．而下三角为少了单位对角元的 L．没有必要储存所有的交换矩阵 I_{ij}．一开始时定义一个向量 v 具有元素$(1,2,3,\cdots,n)$就足够了，而在每一步上简单地对调适当的元素．例如，若第一个主元在第三行中，则$(3,2,1,4,\cdots,n)$可以记录这一点．没有必要具体地将行交换，因此在这个调动中节约了时间．如果需要的话可以从最后的 v 来构造置换矩阵 P．或者就用 v 本身来进行向量 b 之元素的交换．

26.11 将题26.10的方法应用于矩阵

$$A=\begin{bmatrix}0&1&2&3\\3&0&1&2\\2&3&0&1\\1&2&3&0\end{bmatrix}.$$

解 ☞ 基本计算陈列在图26.1中. 在三步之内原始矩阵就被置换成一个4×4的数组包含了除跟踪交换的向量v外所有需要的信息.

在这时矩阵A置换为在PA的LU分解中的一个三角矩阵 向量v告诉我们三角形是显然的, 如果我们以2,3,4,1的顺序来观察行的话. 事实上没有标星号的元素是属于因子U的, L因子也可以通过以同样的行顺序取标有星号的元素来读出. 作为置换矩阵P, 可由在其他地方全为零而在2,3,4,1列各放一个1的方式来构成如下:

$$P=\begin{bmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\\1&0&0&0\end{bmatrix},$$

$$\begin{bmatrix}0&1&2&3\\3&0&1&2\\2&3&0&1\\1&2&3&0\end{bmatrix}$$
已知的矩阵A,
确认第一个主元,3.　　$v=(1,2,3,4)$.

$$\begin{bmatrix}0^*&1&2&3\\ \textcircled{3}&0&1&2\\ \frac{2^*}{3}&3&-\frac{2}{3}&-\frac{1}{3}\\ \frac{1^*}{3}&2&\frac{8}{3}&-\frac{2}{3}\end{bmatrix}$$
将它的行号带到v中的第一个位置上
$v=(2,1,3,4)$.
计算并储存l_{i1}(带星号的).
通过减法计算9个新元素(实线的右侧).

$$\begin{bmatrix}0^*&\frac{1^*}{3}&\frac{20}{9}&\frac{28}{9}\\ 3&0&1&2\\ \frac{2^*}{3}&\textcircled{3}&-\frac{2}{3}&-\frac{1}{3}\\ \frac{1^*}{3}&\frac{2^*}{3}&\frac{28}{9}&-\frac{4}{9}\end{bmatrix}$$
确认第二个主元(2列以及实线的右侧)
将它的行号带至v的第二个位置
$v=(2,3,1,4)$.
计算l_{i2}并储存它们(带星号的).
计算4个新元素.

$$\begin{bmatrix}0^*&\frac{1^*}{3}&\frac{5^*}{7}&\frac{24}{7}\\ 3&0&1&2\\ \frac{2^*}{3}&3&-\frac{2}{3}&-\frac{1}{3}\\ \frac{1^*}{3}&\frac{2^*}{3}&\frac{28}{9}&-\frac{4}{9}\end{bmatrix}$$
确认最后一个主元(3列以及实线的右侧).
将它的行号带至v中第三个位置
$v=(2,3,4,1)$.
计算l_{i3}并储存它们.
计算一个新元素.

图26.1

人们现在可以计算

$$PA=LU=\begin{bmatrix}3&0&1&2\\2&3&0&1\\1&2&3&0\\0&1&2&3\end{bmatrix},$$

并以此来检验采取的所有步骤.

26.12 使用上题的结果并给出 b 向量的分量为(0,1,2,3),解 $Ax=b$.

解 ☞ 我们使用题26.9中的提示.首先不是 Pb 就是向量 v 重新安排 b 的分量成序列(1,2,3,0).虽然这是不必要的,但还是想象我们直接地把方程组 $Ly=Pb$ 列出来.

$$\begin{bmatrix}1&0&0&0\\2/3&1&0&0\\1/3&2/3&1&0\\0&1/3&5/7&1\end{bmatrix}\begin{bmatrix}y_1\\y_2\\y_3\\y_4\end{bmatrix}=\begin{bmatrix}1\\2\\3\\0\end{bmatrix},$$

向前回代则得出 $y=\left(1,\frac{4}{3},\frac{16}{9},-\frac{12}{7}\right)^T$.转向 $Ux=y$ 我们面对

$$\begin{bmatrix}3&0&1&2\\0&3&-2/3&-1/3\\0&0&28/9&-4/9\\0&0&0&24/7\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\\x_4\end{bmatrix}=\begin{bmatrix}1\\4/3\\16/9\\-12/7\end{bmatrix}.$$

由此就来到了 $x=\left(\frac{1}{2},\frac{1}{2},\frac{1}{2},-\frac{1}{2}\right)^T$,它可以在 $Ax=b$ 中直接地得到验证.

26.13 证明线性代数基本定理.

证 ☞ 我们使用 Gauss 算法.假如它可以继续下去到最后,就产生一个三角形方程组,然后通过回代将产生惟一的解,若所有 b_i 为零.这个解有全为零的分量.这已是定理的首要部分.然而假设这个算法不可能继续到预期的三角矩阵结束.这种情况仅当在某一时刻在某根水平线以下的所有系数均为零时才会发生.为明确起见,比如说该算法已经到达了这一时刻.

$$\begin{aligned}u_{11}x_1+\cdots\qquad\qquad&=b'_1,\\u_{22}x_2+\cdots\qquad&=b'_2,\\&\ \vdots\\u_{kk}x_k+\cdots&=b'_k,\\0&=b^{(k)}_{k+1},\\&\ \vdots\\0&=b^{(k)}_n.\end{aligned}$$

于是在齐次的情况下,那里所有的 b' 均为零我们可以随我们所欲选 x_{k+1} 至 x_n,接着决定其他的 x_i.但是在一般情况下,除非 $b^{(k)}_{k+1}$ 至 $b^{(k)}_n$ 都是零,没有解是可能的.若这些 b' 确实发生零的情况,则我们还可以自由地选取 x_{k+1} 至 x_n,之后其他 x_i 被确定.这就是基本定理的内容.

因子分解

26.14 通过对对应元素的直接比较,确定使得 $A=LU$ 成立的 L 和 U 的元素.

解 ☞ 假定无需作交换.这时我们将从让

$$\begin{bmatrix}l&0&0&\cdots&0\\l_{21}&1&0&\cdots&0\\l_{31}&l_{32}&1&\cdots&0\\\cdots&\cdots&\cdots&\cdots&\cdots\\l_{n1}&l_{n2}&l_{n3}&\cdots&1\end{bmatrix}\begin{bmatrix}u_{11}&u_{12}&u_{13}&\cdots&u_{1n}\\0&u_{22}&u_{23}&\cdots&u_{2n}\\0&0&u_{33}&\cdots&u_{3n}\\\cdots&\cdots&\cdots&\cdots&\cdots\\0&0&0&\cdots&u_{nn}\end{bmatrix}=\begin{bmatrix}a_{11}&\cdots&a_{1n}\\a_{21}&\cdots&a_{2n}\\a_{31}&\cdots&a_{3n}\\\cdots&\cdots&\cdots\\a_{n1}&\cdots&a_{nn}\end{bmatrix}$$

两侧的对应元素相等,这相当于 n^2 个未知量 l_{ij} 及 u_{ij} 的 n^2 个方程.行列式的运行如下.首先以 L 的顶行乘 U 的所有列来得到

$$u_{1j}=a_{1j},\qquad j=1,\cdots,n.$$

接着以 L 的行(略去第一行)乘 U 的第一列,得到 $l_{i1}u_{11}=a_{i1}$,由此 l_{i1} 随之而得

$$l_{i1}=\frac{a_{i1}}{u_{11}},\quad i=2,\cdots,n.$$

接下来转到 L 的第二行乘 U 的各列(略去第一列),于是 U 的第二行是

$$u_{2j}=a_{2j}-l_{21}u_{1j},\qquad j=2,\cdots,n,$$

现在以 L 的各行(略去第一、二行)乘 U 的第二列,涉及的所有元素除了 l_{i2} 外已是现成的,所以我们对它求解.

$$l_{i2}=\frac{a_{i2}-l_{i1}u_{12}}{u_{22}},\quad i=3,\cdots,n.$$

以这种递推方式进行下去,我们交替地得到 U 的各行为

$$u_{rj}=a_{rj}-\sum_{k=1}^{r-1}l_{rk}u_{kj},\qquad j=r,\cdots,n.$$

每一行后接着就是 L 相应的列

$$l_{ir}=\frac{a_{ir}-\sum_{k=1}^{r-1}l_{ik}u_{kr}}{u_{rr}},\qquad i=r+1,\cdots,n.$$

这个方法称为 Doolittle 算法.

26.15 什么是 Crout 算法?

解 ☞ Crout 算法也能产生 A 的因子分解,具有形式 $L'U'$,其中 U' 的对角元为 1,L' 有一般的对角元.关于求这些因子的元素,其公式的获得十分类似于题 26.14 中的,但是有意义的是指出,以 D 表示一个对角元与前面 U 的相同,而其他地方均为零的矩阵,便有

$$A=LU=L(DD^{-1})U=(LD)(D^{-1}U)=L'U'.$$

于是二种因子分解是紧密相关联的.

26.16 推出 Choleski 方法用于分解一个实对称正定矩阵.

解 ☞ 这里我们将找到形如

$$A=LL^T$$

的因子,T 表示转置.该过程几乎等同于题 26.14 的那些,由于有对称性允许我们只考虑 A 的下三角.三阶的 Hilbert 矩阵可以再一次充当小型的导引.

$$\begin{bmatrix}l_{11}&0&0\\l_{21}&l_{22}&0\\l_{31}&l_{32}&l_{33}\end{bmatrix}\begin{bmatrix}l_{11}&l_{21}&l_{31}\\0&l_{22}&l_{23}\\0&0&l_{33}\end{bmatrix}=\begin{bmatrix}1&1/2&1/3\\1/2&1/3&1/4\\1/3&1/4&1/5\end{bmatrix}.$$

L 的元素将依照从顶到底和从左到右的顺序得到.

$$l_{11}l_{11}=1,\qquad l_{11}=1.$$

$$l_{21}l_{11}=\frac{1}{2},\qquad l_{21}=\frac{1}{2}.$$

$$l_{21}^2+l_{22}^2=\frac{1}{3},\qquad l_{22}=\frac{1}{\sqrt{12}}.$$

$$l_{31}l_{11}=\frac{1}{3},\qquad l_{31}=\frac{1}{3}.$$

$$l_{31}l_{21}+l_{32}l_{22}=\frac{1}{4},\qquad l_{32}=\frac{1}{\sqrt{12}}.$$

$$l_{31}^2+l_{32}^2+l_{33}^2=\frac{1}{5},\qquad l_{33}=\frac{1}{\sqrt{180}}.$$

计算仍是递推的,每一行只有一个未知量.

因为算法按这种方式展开,我们现在应该有望将我们的努力扩大到对四阶的 Hilbert 矩阵,只需将 L 加边,添上新的底行及第四列.

$$LL^T=\begin{bmatrix}1&0&0&0\\\frac{1}{2}&\frac{1}{\sqrt{12}}&0&0\\\frac{1}{3}&\frac{1}{\sqrt{12}}&\frac{1}{\sqrt{180}}&0\\l_{41}&l_{42}&l_{43}&l_{44}\end{bmatrix}L^T=\begin{bmatrix}1&1/2&1/3&1/4\\1/2&1/3&1/4&1/5\\1/3&1/4&1/5&1/6\\1/4&1/5&1/6&1/7\end{bmatrix}.$$

于是我们得到

$$l_{41}l_{11}=\frac{1}{4},\quad l_{41}=\frac{1}{4}.$$

$$l_{41}l_{21}+l_{42}l_{22}=\frac{1}{5},\quad l_{42}=\frac{3\sqrt{3}}{20}.$$

等等，直至 $l_{43}=\sqrt{5}/20$ 及 $l_{44}=\sqrt{7}/140$.

算法可以用下面的方程来概括

$$\sum_{j=1}^{i-1}l_{rj}l_{ij}+l_{ri}l_{ii}=a_{ri},\quad i=1,\cdots,r-1,$$

$$\sum_{j=1}^{r-1}l_{rj}^2+l_{rr}^2=a_{rr}.$$

依次地用于 $r=1,\cdots,n$.

误差和范数

26.17 什么是矩阵 A 的条件数?

解 ☞ 它是在计算中矩阵可信程度的一个度量.对于一个给定的范数，我们定义条件数为

$$C(A)=\|A\|\cdot\|A^{-1}\|,$$

并看到，使用题 1.34，有 $C(I)=1$，其中 I 是恒等矩阵.此外，使用题 1.38，

$$C(A)=\|A\|\cdot\|A^{-1}\|\geqslant\|I\|=1,$$

所以恒等矩阵有最小的条件数.

26.18 假设方程组 $Ax=b$ 中的向量 b 含有输入误差.估计此类误差对解向量 x 的影响.

解 ☞ 将方程组改写为

$$Ax_e=b+e,$$

与 $Ax=b$ 联合在一起得到

$$A(x_e-x)=e,\qquad x_e-x=A^{-1}e,$$

由它并且使用题 1.60，得出

$$\|x-x_e\|\leqslant\|A^{-1}\|\cdot\|e\|.$$

为把它转变成一个相对误差估计，从 $Ax=b$ 我们有

$$\|A\|\cdot\|x\|\geqslant\|b\|,\qquad \|x\|\geqslant\frac{\|b\|}{\|A\|}.$$

最后

$$\frac{\|x_e-x\|}{\|x\|}\leqslant\|A\|\cdot\|A^{-1}\|\frac{\|e\|}{\|b\|}=C(A)\frac{\|e\|}{\|b\|}.$$

在这个式子中出现条件数.

类似地从

$$\|e\|\leqslant\|A\|\|x_e-x\|\qquad 及\qquad \|A^{-1}\|\cdot\|b\|\geqslant\|x\|,$$

我们得到

$$\frac{\|e\|}{C(A)\|b\|}\leqslant\frac{\|x_e-x\|}{\|x\|},$$

给我们关于相对误差的一个下界和上界.

26.19 假设方程组 $Ax=b$ 的矩阵 A 含有输入误差，估计此类误差对解向量 x 的影响.

解 ☞ 将方程组写成

$$(A+E)x_e=b$$

并与 $Ax=b$ 联合在一起得到

$$A(x_e-x)=-Ex_e,$$

导出

$$\|x_e-x\|\leqslant\|A^{-1}\|\cdot\|E\|\cdot\|x_e\|,$$

$$\frac{\|x_e-x\|}{\|x_e\|}\leqslant\|A^{-1}\|\cdot\|A\|\cdot\frac{\|E\|}{\|A\|}=C(A)\frac{\|E\|}{\|A\|},$$

它估计了相对于解 x_e 的误差.这里又出现 A 的条件数.

这里以及在上题中它度量输入误差被放大了多少.

还可以得到相对于解 x 的估计.此类估计之一为:

$$\frac{\|x_e - x\|}{\|x\|} \leqslant \frac{C(A)(\|E\|/\|A\|)}{1 - C(A)(\|E\|/\|A\|)}.$$

26.20 在计算机只以二位浮点数进行计算的假设下,重新对开局的例子(题 26.1)进行工作.

解 ☞ 现在方程组取形式

$$\begin{aligned} 1.0x_1 + 0.50x_2 + 0.33x_3 &= 1.0,\\ 0.50x_1 + 0.33x_2 + 0.25x_3 &= 0,\\ 0.33x_1 + 0.25x_2 + 0.20x_3 &= 0. \end{aligned}$$

并当 $l_{21}=0.5$ 及 $l_{31}=0.33$ 时它很快就转化为

$$\begin{aligned} 0.08x_2 + 0.09x_3 &= -0.50,\\ 0.09x_2 + 0.09x_3 &= -0.33. \end{aligned}$$

同时第一个方程保持原样.这里我们通过简单的减法计算又可以完成我们所要的三角化.

$$0.01x_2 = 0.17$$

现在向后回代得到 $x_2=17, x_3=-21, x_1=-0.6$ 和一个"解"向量$(-0.6, 17, -21)$.与准确解$(9, -36, 30)$相比,我们看到根本没有任何类似之处.要害是,这方程组的矩阵是声名狼藉的 Hilbert 矩阵家族的一个低等级的成员.将这与我们计算机的严格限制连在一起就导出了一个荒唐的结果.

在题 26.42 中将要得到的逆矩阵为

$$\begin{bmatrix} 9 & -36 & 30\\ -36 & 192 & -180\\ 30 & -180 & 180 \end{bmatrix},$$

在这里的大元素应该受到注意.最大范数是 $36+192+180=408$,使条件数为

$$C(A) = \|A\| \cdot \|A^{-1}\| = \frac{11}{6}(408) = 748.$$

据题 26.19 我们现在有估计式

$$\frac{\|x_e - x\|}{\|x_e\|} \leqslant 748\left(\frac{0.005}{\frac{11}{6}}\right) = 2.04,$$

提示了一个 200% 的相对误差.很明显,前面的计算是幼稚的.至少需要 4 位数字.

作为对比,回顾题 26.6 中如人意的矩阵,它允许即使用一个两位的计算机也能获得精确解.对于那矩阵其最大范数是 2 而逆矩阵也有接近于 2 的范数.于是条件数接近 4 而我们估计为

$$\frac{\|x_e - x\|}{\|x_e\|} \leqslant 4\left(\frac{0.005}{1}\right) = 0.02,$$

或者说 2% 的最大误差.

26.21 什么是"最接近的奇异矩阵"("nearest singular matrix")定理?

解 ☞ 假设 A 是非奇异的而 B 是奇异的.这时,由线代数基本定理,存在着一个向量 $x\neq 0$ 满足 $Bx=0$.对这个 x

$$\|Ax\| = \|Ax - Bx\| = \|(A-B)x\| \leqslant \|A-B\| \cdot \|x\|,$$

并由于 $x = A^{-1}Ax$,我们还可得

$$\|x\| \leqslant \|A^{-1}\| \cdot \|Ax\|.$$

由于 A 是非奇异的,消掉因子 $\|Ax\|$ 便得到

$$\frac{1}{\|A-B\|} \leqslant \|A^{-1}\|,$$

这就是所要求的定理.

它提供的信息是 A 的逆矩阵之大小至少是 A 到最接近的奇异矩阵 B 之"距离"的倒数.

若 A 近乎奇异,则 A^{-1} 会有一个大的范数.若 A 是在 $\|A\|=1$ 的意义下规格化的,条件数还是大的.

作为一个推论我们有下面的直观的结果.

若 B 是"足够靠近"于非奇异阵 A 的,在 $1/\|A-B\|$ 比 $\|A^{-1}\|$ 大的意义下,则 B 也是非奇异的.

26.22 使用题 26.21 的定理来估计在以前的题 1.13 中列出的方程组之矩阵的条件.

$$x_1 + x_2 = 1,$$
$$1.1x_1 + x_2 = 2.$$

解 关键是条件数所需要的 A^{-1} 并不总是容易具有精度地获得.虽然这里不是这种情况,我们观察到系数矩阵接近于奇异矩阵

$$B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

并且得到(使用最大范数)$\|A\| = 2.1$, $\|A-B\| = 0.1$,所以

$$\|A^{-1}\| \geqslant \frac{1}{0.1} = 10, \qquad C(A) \geqslant (2.1)(10) = 21.$$

26.23 估计由于在题 26.22 的第二个方程中用 $1.01x_2$ 来代替 x_2 所造成的误差.

解 误差矩阵为 $E = \begin{bmatrix} 0 & 0 \\ 0 & 0.01 \end{bmatrix}$ 具有最大范数 0.01.因此,我们的估计是

$$\frac{\|x_e - x\|}{\|x_e\|} \leqslant C(A)\frac{\|E\|}{\|A\|} \leqslant 21\left(\frac{0.01}{2.1}\right) = 0.1.$$

对一个 1%的输入误差我们预期一个 10%的输出误差.这种放大是来自 A 的坏条件.正像由 $C(A)$ 所度量的那样.

直接解方程组,我们得到 $x = (10, -9)$ 及 $x_e = (11, -10)$.对于 0.09 的相对误差这使得 $\|x_e - x\| = 1$ 而 $\|x_e\| = 11$.故一个 10%的放大差不难被实现.

26.24 解一个线性方程组的许多中间计算使得舍入误差成为一个重要因素.怎样才能估计这个误差?

解 向后误差分析曾经造就了在这个困难领域中的惟一实实在在的成功.它展示了舍入误差的累积效应可以通过考虑替代方程组 $(A+E)x = b$ 来加以估计,其中 E 是 A 的一个摄动矩阵.接着找寻 E 的元素的界.在 x 中的误差就可以用题 26.19 的公式加以估计.细节远不是那么显然的,然而对大多数解的算法已经进行到底.完整的历史必须在文献中去寻找,但是一种简化了的处理导出部分满意的界

$$\max|e_{ij}| \leqslant n\Delta[\max|a_{ij}| + (3 + n\Delta)\max|b_{ij}|]$$

提供在题 26.113 至 26.117 中.这里的 Δ 依赖于单位舍入误差而 b_{ij} 依赖于已知矩阵 A 的由计算得到的因子 L 及 U.

稍微深刻一些的估计

$$\|E\| \leqslant (1.06\max|u_{ij}|)(3n^2 + n^3)2^{-p}$$

可能应用起来较为方便.例如,若 A 为 10 阶的($n = 10$).而执行计算的位数相当于用 8 位小数($2^{-p} = 10^{-8}$),同时对第一个因子我们作了等于 10 的粗估,则我们得到

$$\|E\| \leqslant (1.3)10^{-4},$$

这提示了或许用来计算的小数位的一半已不再有意义.当然,这个估计是保守的,因为它没有考虑到误差常常会有一定程度的相互抵消.

26.25 系数矩阵 A 的条件数怎样加入舍入误差估计过程?

解 回顾题 26.19.解的相对误差的界为

$$\frac{\|x_e - x\|}{\|x_e\|} \leqslant C(A)\frac{\|E\|}{\|A\|},$$

其中,E 现在是由内部舍入误差引起的对 A 的摄动矩阵.对于一个规格化的 A,x_e 的相对误差因而是两个因子(A 的条件数和 E 的范数)之乘积.

26.26 若有双倍位精度算术可用,舍入误差的情况能有多少改进?

解 据题 26.24 中的公式,若因子 2^{-p} 可以从 10^{-8} 减小至 10^{-16}.会获得外加的 8 位小数无疑是一个重要的改进.然而有一个副作用.对大型方程组而言,即使在单倍位时,要用大量的存储空

间.双倍精度可能正处于两难境地.有一个折衷的办法,类似于人们在题 19.8 中描述过的那样,在那里其着眼点是在计算时间上而不是存储空间.不是在每个计算与存储时都以双倍位进行,而是把这高层次的活动仅限于在算法中随处可见为数众多的内积计算.一旦计算完成,它们的值可以单倍位储存,所造成的舍入误差只是一次而不是本该有的 n 次.在编程时无需费什么力气就能把这个特性结合进去,而其回报能够是可观的.

26.27 近似解 x_e 的残量由向量

$$r = b - Ax_e$$

来定义,它给出线性方程组中每个方程不满足的量.残量与 x_e 的误差之间有何关系?

解 ☞ 由于对精确解有 $Ax = b$,我们有

$$r = A(x - x_e), \qquad x - x_e = A^{-1}r.$$

而且使用题 1.37 有

$$\frac{\|r\|}{\|A\|} \leqslant \|x - x_e\| \leqslant \|A^{-1}\| \cdot \|r\|.$$

由 $Ax = b$ 类似地我们有

$$\|A^{-1}\| \cdot \|b\| \geqslant \|x\| \geqslant \frac{\|b\|}{\|A\|}$$

这样除相应的元素导出所要求的结果,

$$\frac{1}{C(A)} \frac{\|r\|}{\|b\|} \leqslant \frac{\|x - x_e\|}{\|x\|} \leqslant C(A) \frac{\|r\|}{\|b\|}.$$

x_e 的相对误差以相对残量的倍数为其上下界,在这个倍数中包含了 A 的条件数,若 $C(A)$ 接近 1,则相对误差接近相对残量,它当然是毫无困难地适用的.然而,假如 $C(A)$ 是个大数.即使 r 也许是小的数,也有充分理由去怀疑在 x_e 中的不精确性.换言之,假若 A 是坏条件的,方程组可能被一个含有大误差的 x_e 近似地满足.从乐观方面来看,主要看上述不等式的左边,当 $C(A)$ 为大数时,即使残量大还是允许误差为小量,虽然发生这种情况的概率似乎是十分小的.

26.28 什么是迭代改进的方法?

解 ☞ 令 $h = x - x_e$ 并将上题的方程 $A(x - x_e) = r$ 改写为

$$Ah = r,$$

这个方程组有与原始方程相同的系数矩阵.若 A 已被分解,或是 Gauss 消去法的步骤以某种方式被保留,它就以相对小的代价把它解了.手边有了 h,人们计算

$$x = x_e + h.$$

于是得到一个新的并可以推测为更好的对真解的逼近值.现在可以计算新的残量,而只要看来有效果过程就重复下去.这就是迭代改进的思想.假如可以用双倍位算术,这是用它的一个极好的机会.

迭代方法

26.29 使用下面的熟知的例子说明解线性方程组的 Gauss-Seidel 迭代.一条狗遗失在一个方形的多通道迷宫中(图 26.2).在每一个交叉点处它随机地选择一个方向并前进至下一个交叉点,在那儿它又随机地选一个方向等等.一条狗从交叉点 i 处出发最终出现在南面边界上的概率是什么?

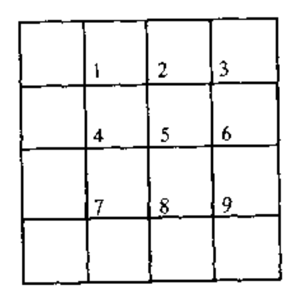


图 26.2

解 ☞ 假定恰好有 9 个内部交叉点,如图所示.令 P_1 代表狗从第 1 个交叉点出发最终出现在南边上的概率,令 $P_2, \cdots, P_9$ 类似地被定义.假设在它到达的每个交叉点,一条狗很可能随意地选一个方向,并假设只要它到达了任何一个出口它的游动就算结束.概率论于是为 P_k 提出下面的 9 个方程:

$$P_1=\frac{1}{4}(0+0+P_2+P_4),\qquad P_2=\frac{1}{4}(0+P_1+P_3+P_5),$$

$$P_3=\frac{1}{4}(0+P_2+0+P_6),\qquad P_4=\frac{1}{4}(P_1+0+P_5+P_7),$$

$$P_5=\frac{1}{4}(P_2+P_4+P_6+P_8)\qquad P_6=\frac{1}{4}(P_3+P_5+0+P_9),$$

$$P_7=\frac{1}{4}(P_4+0+P_8+1),\qquad P_8=\frac{1}{4}(P_5+P_7+P_9+1),$$

$$P_9=\frac{1}{4}(P_6+P_8+0+1).$$

保留方程的这种形式,我们选择 P_k 的初始逼近值.这里有可能作出明智的猜想,但是仍假定对所有的 k 我们取的是平凡的初始值 $P_k=0$.按列出的顺序取方程,我们逐个地计算第二次的逼近值.首先 P_1 算出来为零.接着 $P_2,\cdots,P_6$ 也如此.但是这之后我们得到

$$P_7=\frac{1}{4}(0+0+0+1)=\frac{1}{4},\qquad P_8=\frac{1}{4}\left(0+\frac{1}{4}+0+1\right)=\frac{5}{16},$$

$$P_9=\frac{1}{4}\left(0+\frac{5}{16}+0+1\right)=\frac{21}{64},$$

这样每个 P_k 的第二次逼近值就有了.注意在计算 P_8 和 P_9 时分别用到 P_7 和 P_8 的最新逼近值.看来使用早些时的逼近值意义不大.这个(使用最新值的)过程更快地导出正确结果.现在以同样的方式逐次得到逼近值,迭代持续到在所要求的小数位不出现进一步的变化为止.工作到 3 位,得到表 26.1 中的结果.注意算出的 P_5 为 0.250,这意味着从中心出发的狗之中有 1/4 会出现在南面的边界上.从对称性的角度,这是合理的.可以将所有的 9 个值回代入原始方程作进一步检验,来看诸残量是否是小量.

表 26.1

迭代	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0.250	0.312	0.328
2	0	0	0	0.062	0.078	0.082	0.328	0.394	0.328
3	0.016	0.024	0.027	0.106	0.152	0.127	0.375	0.464	0.398
4	0.032	0.053	0.045	0.140	0.196	0.160	0.401	0.499	0.415
5	0.048	0.072	0.058	0.161	0.223	0.174	0.415	0.513	0.422
6	0.058	0.085	0.065	0.174	0.236	0.181	0.422	0.520	0.425
7	0.065	0.092	0.068	0.181	0.244	0.184	0.425	0.524	0.427
8	0.068	0.095	0.070	0.184	0.247	0.186	0.427	0.525	0.428
9	0.070	0.097	0.071	0.186	0.249	0.187	0.428	0.526	0.428
10	0.071	0.098	0.071	0.187	0.250	0.187	0.428	0.526	0.428

在 Gauss-Seidel 方法的这一例子里 9 个方程中的每一个都以下面的形式出现:

$$P_i=\cdots,$$

并且以此形式来校正对 P_i 的逼近,用到的是其他分量的最新值.值得注意的是在每个方程里左边的那个未知量有最大的系数.

26.30 对一个一般的线性方程组推出 Gauss-Seidel 方法.

解 ☞ 该算法最经常的是应用于具有对角占优的矩阵 A 的方程组 $Ax=b$.在任何情况下,人们应该通过行和列的对调进行整理,尽可能地使较大的元素落在对角线上.方程组的第 i 个方程将 x_i 用其他的未知量来表示而解出它.假如我们使用记号 $x_i^{(k)}$ 表对 x_i 的第 k 次逼近,则算法的进行如例子中那样.

$$x_1^{(1)}=\frac{b_1-a_{12}x_2^{(0)}-\cdots-a_{1n}x_n^{(0)}}{a_{11}},$$

$$x_2^{(1)} = \frac{b_2 - a_{21}x_1^{(1)} - a_{23}x_3^{(0)} - \cdots - a_{2n}x_n^{(0)}}{a_{22}},$$

$$x_3^{(1)} = \frac{b_3 - a_{31}x_1^{(1)} - a_{32}x_2^{(1)} - a_{34}x_4^{(0)} - \cdots - a_{3n}x_n^{(0)}}{a_{33}},$$

$$\vdots$$

上标(0)表初始逼近.更一般地我们对 x_i 的 k 次逼近有

$$x_i^{(k)} = \frac{b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)}}{a_{ii}},$$

在这里面第一个和使用了对具有 $j<i$ 的 x_j 的第 k 次逼近,而第二项和使用了具有 $j>i$ 的 x_j 的第 $(k-1)$ 次逼近.此处 $i=1,\cdots,n$ 和 $k=1,\cdots$.

26.31 表示 Gauss-Seidel 算法为矩阵形式.

解 ☞ 首先矩阵 A 被分裂成

$$A = L + D + U,$$

其中 L 及 U 是对角线上元素为零的下三角与上三角阵.于是题 26.30 的一般公式可以写成

$$x^{(k)} = D^{-1}(b - Lx^{(k)} - Ux^{(k-1)})$$

它可以就 $x^{(k)}$ 解出.首先

$$(I + D^{-1}L)x^{(k)} = D^{-1}b - D^{-1}Ux^{(k-1)},$$

它导出

$$x^{(k)} = (I + D^{-1}L)^{-1}(D^{-1}b - D^{-1}Ux^{(k-1)}),$$

或是

$$x^{(k)} = -(D+L)^{-1}Ux^{(k-1)} + (D+L)^{-1}b.$$

26.32 什么是一个定常的矩阵迭代?

解 ☞ 一个具有形式

$$x^{(k)} = M_k x^{(k-1)} + C_k b$$

的矩阵迭代称为定常的,若 M_k 与 C_k 不依赖于 k.于是迭代变成

$$x^{(k)} = Mx^{(k-1)} + Cb.$$

Gauss-Seidel 方法是定常的,它具有这种 M 及 C.

$$M = -(D+L)^{-1}U \qquad C = (D+L)^{-1}.$$

26.33 讨论矩阵迭代的收敛性.

解 ☞ 首先我们要求 $Ax=b$ 的精确解是迭代法的一个不动点.即,我们将 $x=A^{-1}b$ 来替代

$$x^{(k)} = M_k x^{(k-1)} + C_k b$$

中的输入逼近值与输出逼近值,因而有

$$x = A^{-1}b = M_k A^{-1}b + C_k b = M_k x + C_k b,$$

它对所有向量 b 均成立,所以我们令系数相等.

$$A^{-1} = M_k A^{-1} + C_k,$$
$$I = M_k + C_k A.$$

现在我们定义 $e^{(k)}$ 为 k 次逼近值的误差.

$$e^{(k)} = x - x^{(k)},$$

于是

$$e^{(k)} = x - M_k x^{(k-1)} - C_k b$$
$$= M_k(x - x^{(k-1)}) = M_k e^{(k-1)}.$$

这表明控制误差性态的就是 M_k.重复使用这个结果,

$$e^{(k)} = M_k M_{k-1} \cdots M_1 e^{(0)},$$

其中 $e^{(0)}$ 是初始误差,对一个定常的迭代而言它就成了

$$e^{(k)} = M^k e^{(0)}.$$

26.34 证明:假如矩阵 A 是正定、对称的,则 Gauss-Seidel 迭代对任意的一个初始向量 $x^{(0)}$ 收敛.

证 ☞ 因为对称性,$A=L+D+L^T$,它使

$$M=-(D+L)^{-1}L^T.$$

若 λ 及 v 是 M 的一个特征值和特征向量,则

$$(D+L)^{-1}L^Tv=-\lambda v,$$
$$L^Tv=-\lambda(D+L)v.$$

以 v 的共轭转置(以 v^* 来表示)来预乘,得

$$v^*L^Tv=-v^*\lambda(D+L)v,$$

然后将 $v^*(D+L)v$ 加到两边

$$v^*Av=(1-\lambda)v^*(D+L)v,$$

因为 $A=L+D+L^T$. 然而 v^*Av 的共轭转置为 v^*Av, 所以对于上个方程的右边同样必定如此. 因此以 $\bar{\lambda}$ 表 λ 的共轭,有

$$(1-\bar{\lambda})v^*(D+L)^Tv=(1-\lambda)v^*(D+L)v=(1-\lambda)(v^*Dv+v^*Lv)$$
$$=(1-\lambda)(v^*Dv-\bar{\lambda}v^*(D+L)^Tv).$$

将项进行合并,

$$(1-|\lambda|^2)v^*(D+L)^Tv=(1-\lambda)v^*Dv.$$

再将两边乘以 $(1-\bar{\lambda})$,并作少量代数运算最后我们有

$$(1-|\lambda|^2)v^*Av=|1-\lambda|^2v^*Dv.$$

但是 v^*Av 及 v^*Dv 都是非负的因而 λ 不可能等于 1(因为这会带回到 $Av=0$),所以

$$|\lambda|^2<1.$$

所有特征值都位于一个单位圆内并保证 $\lim M^k=0$. 因此对任何 $e^{(0)}$ 来说 $e^{(k)}$ 的极限为零.

26.35 怎样才能将一个加速方法应用于 Gauss-Seidel 迭代?

解 ☞ 由于 $e^{(k)}=Me^{(k-1)}$, 我们预计误差可以按一个常数速率下降,十分像在题 25.4 那样. 外推到极限的思想也就使人联想起它本身. 这里它该取的形式为

$$x_i=x_i^{(k+2)}-\frac{\Delta x_i^{(k+1)}}{\Delta^2x_i^{(k)}},$$

对 $i=1,\cdots,n$. 上标表示相继的三个逼近值.

例如,使用表 26.1 的中心列,在该列中我们知道校正值为 0.250, 在第 4 行至第 8 行的误差为在第三位小数上的 54, 27, 14, 6, 与 3. 这非常接近于稳定地减半. 假设我们尝试外推到极限使用的是下面的三个元素,与它们连同对应的差分一起如下所给出.

0.196		
	0.027	
0.223		−0.014
	0.013	
0.236		

我们得到

$$P_5=0.236-\frac{(0.013)^2}{-0.014}=0.248$$

它是在正确的方向上,假如不是特别戏剧性的话.

26.36 什么是松弛法与超松弛法?

解 ☞ 中心思想是使用残量作为如何校正已经得到的逼近值的指示器,例如迭代公式

$$x^{(k)}=x^{(k-1)}+(b-Ax^{(k-1)})$$

有松弛法的特征. 已经发现给予残量一个外加的权可以加速收敛, 由此导出超松弛公式诸如

$$x^{(k)}=x^{(k-1)}+w(b-Ax^{(k-1)}),$$

取 $w>1$. 这个思想的其他变化也已经被使用过.

26.37 改编超松弛法来加速 Gauss-Seidel 收敛.

解 ☞ 自然的改编为

$$x^{(k)} = x^{(k-1)} + w[b - Lx^{(k)} - (D+U)x^{(k-1)}],$$

具有 $A = L + D + U$ 如前. 我们取 $w = 1.2, x^{(0)} = 0$, 并再一次尝试去解狗在迷宫中的问题. 我们发现生成的零如前直至

$$P_7^{(1)} = P_7^{(0)} + 1.2\left(0.250 + \frac{1}{4}P_4^{(1)} - P_7^{(0)} + \frac{1}{4}P_8^{(0)}\right) = 0.300,$$

$$P_8^{(1)} = P_8^{(0)} + 1.2\left(0.250 + \frac{1}{4}P_5^{(1)} + \frac{1}{4}P_7^{(1)} - P_8^{(0)} + \frac{1}{4}P_9^{(0)}\right) = 0.390,$$

$$P_9^{(1)} = P_9^{(0)} + 1.2\left(0.250 + \frac{1}{4}P_6^{(1)} + \frac{1}{4}P_8^{(1)} - P_9^{(0)}\right) = 0.418.$$

以同样的方式得到逐次逼近值并列在表 26.2 中. 注意现在所需迭代次数约为一半.

表 26.2

Iteration	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0.300	0.390	0.418
2	0	0	0	0.090	0.144	0.169	0.384	0.506	0.419
3	0.028	0.052	0.066	0.149	0.234	0.182	0.420	0.520	0.427
4	0.054	0.096	0.071	0.183	0.247	0.187	0.427	0.526	0.428
5	0.073	0.098	0.071	0.188	0.251	0.187	0.428	0.527	0.428
6	0.071	0.098	0.071	0.187	0.250	0.187	0.428	0.526	0.428

矩阵求逆

26.38 推广 Gauss 消去法来产生系数矩阵 A 的逆矩阵. 即, 求 A^{-1} 使 $AA^{-1} = I$.

解 ☞ 再一次用题 26.1 的方程组, 我们简单地同时处理三个 b 向量. 从数组

$$\begin{matrix} 1 & \frac{1}{2} & \frac{1}{3} & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & 0 & 1 & 0 \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & 0 & 0 & 1 \end{matrix}$$

出发, 左半边是 A 而右半边是 I. 现在第一个 Gauss 步导出的新数组为

$$\begin{matrix} 1 & \frac{1}{2} & \frac{1}{3} & 1 & 0 & 0 \\ 0 & \frac{1}{12} & \frac{1}{12} & -\frac{1}{2} & 1 & 0 \\ 0 & \frac{1}{12} & \frac{4}{45} & -\frac{1}{3} & 0 & 1 \end{matrix}$$

这里让第二个主元还原为 1 将方法稍加改变, 通过将第二行乘以 12 来执行这一措施.

$$\begin{matrix} 1 & \frac{1}{2} & \frac{1}{3} & 1 & 0 & 0 \\ 0 & 1 & 1 & -6 & 12 & 0 \\ 0 & 0 & \frac{1}{180} & \frac{1}{6} & -1 & 1 \end{matrix}$$

第二步还完成了方程组的三角化. 此时可以使用向后回代来解三个不同的方程组, 每一组包含最后三列向量中的一列. 然而, 我们不去回代而是扩充第二 Gauss 步. 继续以第二行为主元行, 我们从第一行中减去它的一半为了多生成一个零,

$$\begin{matrix} 1 & 0 & 1/6 & 4 & 6 & 0 \\ 0 & 1 & 1 & 6 & 12 & 0 \\ 0 & 0 & 1/180 & 1/6 & -1 & 1 \end{matrix}$$

将最后一个主元还元为 1 后，接着第三 Gauss 步．这一步的目的是将新的主元上方生成零．于是就出现最后的数组．

$$\begin{matrix} 1 & 0 & 0 & 9 & -36 & 30 \\ 0 & 1 & 0 & -36 & 192 & -180 \\ 0 & 0 & 1 & 30 & -180 & 180 \end{matrix}$$

由于我们实际上已经解了三个形如 $Ax=b$ 的方程组，依次带有向量 $b^T=(1,0,0)$，$(0,1,0)$及$(0,0,1)$，很清楚，最后三列现在包含 A^{-1}．原始的数组是(A,I)．最后的数组是(I,A^{-1})．对其他的矩阵 A 也可以应用同样的过程，倘若需要的话可作行或列的交换．假如作了这种交换，在算法完成时必须予以恢复．

26.39　假设矩阵 A 已被因子分解成$A=LU$，怎样能从因子中求得 A^{-1}?

解　☞ 由于 $A^{-1}=U^{-1}L^{-1}$，就化成了三角矩阵求逆的问题．考虑 L 并寻找一个同样形式的逆矩阵．

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & l_{32} & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ c_{21} & 1 & 0 & \cdots & 0 \\ c_{31} & c_{32} & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & 1 \end{bmatrix} = LL^{-1} = I,$$

当我们在进行时便可清楚这个假设的合法性．现在令二边相应的元素相等，十分像在 Choleski 因子分解算法中的那样，从上到下从左到右．我们得到

$$\begin{aligned} l_{21}+c_{21}=0, &\quad c_{21}=-l_{21}, \\ l_{31}+l_{32}c_{21}+c_{31}=0, &\quad c_{31}=-(l_{31}+l_{32}c_{21}), \\ l_{32}+c_{32}=0, &\quad c_{32}=-l_{32}, \\ l_{41}+l_{42}c_{21}+l_{43}c_{31}+c_{41}=0, &\quad c_{41}=-(l_{41}+l_{42}c_{21}+l_{43}c_{31}), \\ l_{42}+l_{43}c_{32}+c_{42}=0, &\quad c_{42}=-(l_{42}+l_{43}c_{32}), \\ l_{43}+c_{43}=0, &\quad c_{43}=-l_{43}, \\ &\vdots \end{aligned}$$

元素递推地确定．一般的公式是

$$c_{ij}=-\sum_{k=j}^{i-1}l_{ik}c_{kj},\qquad \begin{matrix} i=2,\cdots,n, \\ j=1,\cdots,i-1. \end{matrix}$$

所有对角元均为 1．

U 的求逆类似．假设逆矩阵是上三角阵，具有元素 d_{ij}，我们从下到上从右到左，得到．

$$d_{ii}=\frac{1}{u_{ii}}\quad i=n,\cdots,1$$

及

$$d_{ij}=\frac{-1}{u_{ii}}\sum_{k=i+1}^{j}u_{ik}d_{kj},\qquad \begin{matrix} i=n,\cdots,1, \\ j=n,\cdots,i+1. \end{matrix}$$

26.40　应用上题的方法于题 26.11 的矩阵．

解　☞ 在那个问题中分解的因子为

$$PA=LU=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 2/3 & 1 & 0 & 0 \\ 1/3 & 2/3 & 1 & 0 \\ 0 & 1/3 & 5/7 & 1 \end{bmatrix}\begin{bmatrix} 3 & 0 & 1 & 2 \\ 0 & 3 & -2/3 & -1/3 \\ 0 & 0 & 28/9 & -4/9 \\ 0 & 0 & 0 & 24/7 \end{bmatrix},$$

应用上面的递推公式，我们现在有

$$L^{-1}=\frac{1}{63}\begin{bmatrix}63&0&0&0\\42&63&0&0\\7&-42&63&0\\9&9&-45&63\end{bmatrix},$$

$$U^{-1}=\frac{1}{168}\begin{bmatrix}56&0&-18&-35\\0&56&12&7\\0&0&54&7\\0&0&0&49\end{bmatrix},$$

由它们最后得出

$$(PA)^{-1}=U^{-1}L^{-1}=\frac{1}{24}\begin{bmatrix}7&1&1&-5\\-5&7&1&1\\1&-5&7&1\\1&1&-5&7\end{bmatrix},$$

为了产生最终的 A^{-1}，我们使用 $A^{-1}=(PA)^{-1}P$ 并回忆以一个置换矩阵 P 后乘就等于将列重新安排. 返回去参看前面的问题,发现上述诸列应取成 4,1,2,3 的顺序.

26.41 推导关于在一个线性方程组中作一个交换步的公式.

解 ☞ 令线性方程组为 $Ax=b$，或是

$$\sum_{k=1}^{n}a_{ik}x_k=b_i,\qquad i=1,\cdots,n,$$

它基本的组成可以用一个 $n=3$ 的数组加以说明.

	x_1	x_2	x_3
b_1	a_{11}	a_{12}	a_{13}
b_2	a_{21}	a_{22}	a_{23}
b_3	a_{31}	a_{32}	a_{33}

我们进行将一个“因”(“dependent”)变量(譬如,b_2)与一个自(independent)变量(譬如,x_3)的交换. 从第二个方程中解出 x_3，$x_3=(b_2-a_{21}x_1-a_{22}x_2)/a_{23}$. 这要求主元系数 a_{23} 不为零,将 x_3 的这个表达式代到余下的二个方程中去便带来

$$b_1=a_{11}x_1+a_{12}x_2+\frac{a_{13}(b_2-a_{21}x_1-a_{22}x_2)}{a_{23}},$$

$$b_3=a_{31}x_1+a_{32}x_2+\frac{a_{33}(b_2-a_{21}x_1-a_{22}x_2)}{a_{23}}.$$

在对换之后,新方程组的数组如下:

	x_1	x_2	b_2
b_1	$a_{11}-\frac{a_{13}a_{21}}{a_{23}}$	$a_{12}-\frac{a_{13}a_{22}}{a_{23}}$	$\frac{a_{13}}{a_{23}}$
x_3	$-\frac{a_{21}}{a_{23}}$	$-\frac{a_{22}}{a_{23}}$	$\frac{1}{a_{23}}$
b_3	$a_{31}-\frac{a_{33}a_{21}}{a_{23}}$	$a_{32}-\frac{a_{33}a_{22}}{a_{23}}$	$\frac{a_{33}}{a_{23}}$

它可以总结成四条规则:

1. 主元系数置换成它的倒数.
2. 主元列的剩余部分用主元系数去除.
3. 主元列剩余部分以变号的主元系数去除.
4. 任何其他系数(比如 a_{lm})被置换成 $a_{lm}-\frac{a_{lk}a_{jm}}{a_{jk}}$ 其中 a_{jk} 为主元系数.

26.42 以例子说明用于寻找逆矩阵的交换法.

解 ☞ 我们再一次地取题 26.1 的矩阵.

	x_1	x_2	x_3
b_1	1	1/2	1/3
b_2	1/2	1/3	1/4
b_3	1/3	1/4	1/5

为控制误差,习惯的做法是,选择最大系数来作主元,在目前情况下选 1. 将 b_1 与 x_1 对调,我们得到新的数组

	b_1	x_2	x_3
x_1	1	$-\frac{1}{2}$	$-\frac{1}{3}$
b_2	$\frac{1}{2}$	$\frac{1}{12}$	$\frac{1}{12}$
b_3	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{4}{45}$

二个类似的对换:b_3 与 x_3,b_2 与 x_2 导出下面所示的二个数组. 在每种情况下在一 b 行与 x 列中的最大系数用作主元.

	b_1	x_2	b_3
x_1	9/4	-3/16	-15/4
b_2	3/16	1/192	15/16
x_3	-15/4	-15/16	45/4

	b_1	b_2	b_3
x_1	9	-36	30
x_2	-36	192	-180
x_3	30	-180	180

由于我们所做的是将方程组 $b = Ax$ 换成方程组 $x = A^{-1}b$,所以最后的矩阵就是 A^{-1}.

26.43 导出公式 $A^{-1} = (I + R + R^2 + \cdots)B$,其中

$$R = I - BA.$$

解 ☞ 这里的想法是 B 为 A 的一个近似逆,所以残量 R 有小的元素,因而级数中取少数几项可能足以产生一个好得多的对 A^{-1} 近似. 为了导出这个公式,首先指出 $(I - R)(I + R + R^2 + \cdots) = I$ 是以矩阵级数收敛为条件的. 这时 $I + R + R^2 + \cdots = (I - R)^{-1}$ 并因此

$$(I + R + R^2 + \cdots)B = (I - R)^{-1}B = (BA)^{-1}B$$
$$= A^{-1}B^{-1}B,$$

它还原到 A^{-1}.

26.44 应用上题的公式于矩阵

$$A = \begin{bmatrix} 1 & 10 & 1 \\ 2 & 0 & 1 \\ 3 & 3 & 2 \end{bmatrix},$$

假设只有一个 3-位计算机可用. 由于任何计算机都只能进行有限位数的计算,所以这将再一次说明逐次校正方法的能力.

解 ☞ 首先我们应用 Gauss 消去法得到这个逆矩阵的首次近似. 下面列出了消去法的三步,在每一步都使用恰当的最大主元,近似逆矩阵 B 也一起列出,它是经过两次行交换把底行移到顶行而得来的.

$$\begin{array}{cccccc} 0.1 & 1 & 0.1 & 0.1 & 0 & 0 \\ 2.0 & 0 & 1.0 & 0 & 1 & 0 \\ 2.7 & 0 & 1.7 & -0.3 & 0 & 1 \end{array} \qquad \begin{array}{cccccc} 0 & 1 & 0.037 & 0.111 & 0 & -0.0371 \\ 0 & 0 & -0.260 & 0.222 & 1 & -0.742 \\ 1 & 0 & 0.630 & -0.111 & 0 & 0.371 \end{array}$$

第一步　　　　第二步

$$\begin{array}{cccccc} 0 & 1 & 0 & 0.143 & 0.143 & -0.143 \\ 0 & 0 & 1 & 0.854 & -3.85 & 2.85 \\ 1 & 0 & 0 & 0.427 & 2.43 & -1.43 \end{array} \qquad \begin{bmatrix} 0.427 & 2.43 & -1.43 \\ 0.143 & 0.143 & -0.143 \\ 0.854 & 3.85 & 2.85 \end{bmatrix}$$

第三步　　　　矩阵 B

接下来我们方便地计算

$$R = I - BA = \begin{bmatrix} 0.003 & 0.020 & 0.003 \\ 0 & -0.001 & 0 \\ 0.004 & -0.010 & 0.004 \end{bmatrix}.$$

在这之后按下面的顺序求 RB, $B+RB$, $R^2B-R(RB)$, 及 $B+RB+R^2B$ (注意由于 R^2B 是如此小, 为了表示的简单起见引进了一个 10,000 的因子.)

$$\begin{bmatrix} 0.001580 & -0.001400 & 0.001400 \\ -0.000143 & -0.000143 & 0.000143 \\ -0.003140 & -0.007110 & 0.007110 \end{bmatrix} \qquad \begin{bmatrix} 0.428579 & 2.428600 & -1.428600 \\ 0.142857 & 0.142857 & -0.142857 \\ -0.857138 & -3.857110 & 2.857110 \end{bmatrix}$$

RB　　　　$B+RB$

$$\begin{bmatrix} -0.07540 & -0.28400 & 0.28400 \\ 0.00143 & 0.00143 & -0.00143 \\ -0.04810 & -0.32600 & 0.32600 \end{bmatrix} \qquad \begin{bmatrix} 0.4285715 & 2.4285716 & -1.4285716 \\ 0.1428571 & 0.1428571 & -0.1428571 \\ -0.8571428 & -3.8571426 & 2.8571426 \end{bmatrix}$$

$10^4 \cdot R(RB)$　　　　$B+RB+R^2B$

注意除了在加法过程中,都是以 3 位有效数字进行计算.由于精确逆矩阵为

$$A^{-1} = \frac{1}{7}\begin{bmatrix} 3 & 17 & -10 \\ 1 & 1 & -1 \\ -6 & -27 & 20 \end{bmatrix}.$$

可以验证 $B+RB+R^2B$ 只在第 7 位小数上有误差.取级数公式的更多项将会带来还要进一步的精度.此法常用于改进由 Gauss 消去法求逆的结果,因为 Gauss 方法对舍入误差的积累过于敏感.

行列式

26.45 行列式不再广泛地使用于解线性方程,但继续有其他方面的应用,直接计算一个 n 阶行列式会要求 $n!$ 项的计算,它是禁用的,除非对小的 n.别的选择是什么?

解 ☞ 从行列式的性质,Gauss 消去法中没有哪一步能改变矩阵的行列式的值,规格化及对换除外.假如没有执行这些运算,行列式值可借助三角化后诸对角元相乘而得.因此,对于题 26.1 的矩阵其行列式值可作快速计算:$\left(\frac{1}{12}\right)\left(\frac{1}{180}\right)=\frac{1}{2160}$.这值之小是矩阵的麻烦特性的另外一个象征.

行列式值还可以从因子分解 $PA=LU$ 得到.

由于 $A=P^{-1}LU$ 我们有

$$\det(A) = \det(P^{-1})\det(L)\det(U) = (-1)^p\det(U),$$

其中 p 是由置换矩阵 P 或 P^{-1} 所表示的交换之次数.对于题 26.11 的矩阵

$$\det(U) = 3(3)\left(\frac{28}{9}\right)\left(\frac{24}{7}\right) = 96$$

而 $\det(P)$ 易见其值为 -1.(或者回顾在分解因子时作了三次对换,使 $p=3$.)因此

$$\det(A) = -96.$$

特征值问题,特征多项式

26.46 什么是矩阵 A 的特征值与特征向量?

解 ☞ 一个数 λ，对它来说方程组 $Ax=\lambda x$ 或者 $(A-\lambda I)x=0$ 有一个非零解向量 x，该数就称为方程组的特征值，任何对应的非零解向量 x 称为一个特征向量．显然，若 x 是一个特征向量则对任何数 C，Cx 也是．

26.47 寻找方程组

$$\begin{aligned}(2-\lambda)x_1-x_2&=0,\\ -x_1+(2-\lambda)x_2-x_3&=0,\\ -x_2+(2-\lambda)x_3&=0.\end{aligned}$$

的特征值与特征向量．在包括由三质点以弹簧连结在一起组成的系统之振动在内的各种物理装置中会出现这类方程组．

解 ☞ 我们以例说明，直接地寻找特征多项式并接着获得后作为这个多项式之零点的特征值的方法．最后得到特征向量．这第一步是取方程的线性组合，很像在 Gauss 消去法中那样，直到只有 x_3 列的系数中包含 λ．例如若以 E_1、E_2，E_3 表示这三个方程，则 $-E_2+\lambda E_3$ 是方程

$$x_1-2x_2+(1+2\lambda-\lambda^2)x_3=0,$$

称它为 E_4，组合 $E_1-2E_2+\lambda E_4$ 变成

$$4x_1-5x_2+(2+\lambda+2\lambda^2-\lambda^3)x_3=0,$$

这最后的两个方程与 E_3 一起，现在只有 x_3 的系数中包含 λ．

过程的第二步是将方程组以 Gauss 消去法或者等价的方法进行三角化．以这个小方程组对待主元我们可以有少许的自由，保留

$$\begin{aligned}x_1-2x_2+(1+2\lambda-\lambda^2)x_3&=0,\\ -x_2+(2-\lambda)x_3&=0\end{aligned}$$

作为我们的前二个方程并立刻实现

$$(4-10\lambda+6\lambda^2-\lambda^3)x_3=0$$

来完成三角化．为了满足这最后一个方程我们必须避免使 $x_3=0$，因为它立刻迫使 $x_2=x_1=0$，从而我们就没有一个非零解向量．据此我们必须要求

$$4-10\lambda+6\lambda^2-\lambda^3=0.$$

这个三次式是特征多项式，而特征值必定是它的零点，因为没有其他的方法可以得到一个非零解向量．据早前章节中的方法我们得到这些特征值按递增的顺序为 $\lambda_1=2-\sqrt{2}$，$\lambda_2=2$，$\lambda_3=2+\sqrt{2}$．

最后一步是找特征向量，然而，对于已经三角化了的方程组，只要用向后回代就行了．首先取 λ_1，并记住特征向量只要被决定到含有任意的乘数，所以我们可以选 $x_3=1$，我们得到 $x_2=\sqrt{2}$ 接着 $x_1=1$．其他的特征向量使用 λ_2 与 λ_3 以同样的方法得到．

最后的结果是

λ	x_1	x_2	x_3
$2-\sqrt{2}$	1	$\sqrt{2}$	1
2	-1	0	1
$2+\sqrt{2}$	1	$-\sqrt{2}$	1

在本题情况下三个方程的原始方程组有三个不同的特征值，对于它们中的每一个都对应了一个独立的特征向量．这是最简单的，然而并不是惟一可能出现的特征值问题．应该指出，现在的这个矩阵既是实的还是对称的．对于一个实对称的 $n\times n$ 矩阵，代数学的一个重要定理陈述了：

(a)所有特征值均为实的，虽然或许不是不同的；

(b)总存在有 n 个独立的特征向量．

这并不是对所有矩阵都成立的．幸运的是计算机目前面临的许多矩阵问题是既实又对称的．

26.48 为了使直接计算特征多项式算法更为清楚，将它应用到这个更大的方程组：

$$\begin{aligned}
E_1:&\quad (1-\lambda)x_1 + x_2 + x_3 + x_4 = 0,\\
E_2:&\quad x_1 + (2-\lambda)x_2 + 3x_3 + 4x_4 = 0,\\
E_3:&\quad x_1 + 3x_2 + (6-\lambda)x_3 + 10x_4 = 0,\\
E_4:&\quad x_1 + 4x_2 + 10x_3 + (20-\lambda)x_4 = 0.
\end{aligned}$$

解 ☞ 记这些方程为 E_1,E_2,E_3,E_4，组合 $E_1+4E_2+10E_3+\lambda E_4$ 为

$$15x_1+39x_2+73x_3+(117+20\lambda-\lambda^2)x_4=0,$$

它是我们的第二个方程，在它里面除了 x_4 项外都与 λ 无关，我们立刻开始通过减 $15E_4$ 进行三角化得到

$$E_5:\quad -21x_2-77x_3+(-183+35\lambda-\lambda^2)x_4=0.$$

组合 $-21E_2-77E_3+\lambda E_5$ 变成

$$-98x_1-273x_2-525x_3+(-854-183\lambda+35\lambda^2-\lambda^3)x_4=0,$$

它是我们的第三个方程，在它里面所有项除 x_4 项外均与 λ 无关．继续三角化通过联结这最后的方程与 E_4 及 E_5 得到

$$E_6:\quad 392x_3+(1449-1736\lambda+616\lambda^2-21\lambda^3)x_4=0.$$

现在形成 $392E_3+\lambda E_6$ 的组合，

$$392x_1+1176x_2+2352x_3+(3920+1449\lambda-1736\lambda^2+616\lambda^3-21\lambda^4)x_4=0.$$

而三角化的完成靠联结这个方程与 E_4,E_5 及 E_6 得到

$$E_7:\quad (1-29\lambda+72\lambda^2-29\lambda^3+\lambda^4)x_4=0.$$

现在方程组 E_4,E_5,E_6,E_7 就是我们所瞄准的三角方程组．为了避免零解向量，λ 必须是作为特征多项式的 $1-29\lambda+72\lambda^2-29\lambda^3+\lambda^4$ 的一个零点．它就寻找这些零点及相应的特征向量留作一个练习．刚才所使用的程序可以推广到更大的方程组．

26.49 以例说明在寻找矩阵的特征方程时 Cayley-Hamilton 定理的用途．

解 ☞ 将方程写成

$$f(\lambda)=\lambda^n+c_1\lambda^{n-1}+\cdots+c_{n-1}\lambda+c_n=0,$$

Cayley-Hamilton 定理阐明矩阵 A 本身也满足这个方程．即

$$f(A)=A^n+c_1A^{n-1}+\cdots+c_{n-1}A+c_nI=0,$$

其中在右侧的现在是零矩阵，这就出现关于 n 个系数 c_i 的 n^2 个方程，因而有实质性的冗余．

例如，取 Fibonacci 矩阵 $F=\begin{bmatrix}1&1\\1&0\end{bmatrix}$，由于 $F^2=\begin{bmatrix}2&1\\1&1\end{bmatrix}$，我们有

$$\begin{bmatrix}2&1\\1&1\end{bmatrix}+c_1\begin{bmatrix}1&1\\1&0\end{bmatrix}+c_2\begin{bmatrix}1&0\\0&1\end{bmatrix}=\begin{bmatrix}0&0\\0&0\end{bmatrix}$$

或

$$2+c_1+c_2=0,$$

$$1+c_1=0,\qquad 1+c_2=0,$$

它们中的第二个应重复一遍．熟悉的方程 $\lambda^2=\lambda+1$ 就又在手边．(参看题 18.24 及 26.128.)

或者考虑置换矩阵 P 具有

$$P=\begin{bmatrix}0&0&1\\1&0&0\\0&1&0\end{bmatrix},\quad P^2=\begin{bmatrix}0&1&0\\0&0&1\\1&0&0\end{bmatrix},\quad P^3=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}.$$

它很快地引出方程组

$$1+c_3=0,\qquad c_1=0,\qquad c_2=0,$$

重复二次．特征方程为 $\lambda^3-1=0$．

好几种方案曾经被提出用来在这 n^2 个可用的方程中选取一个适当的子集．方案之一需要计算

$$f(A)v=0.$$

对一个适当的向量 v，并解这个方程组．

26.50 证明 Gerschgorin 定理，它阐明矩阵 A 的每一个特征值都落在一个复圆内，它们分别以 a_{ii} 为中心，以，

$$R_i = \sum_{j\neq i} a_{ij} \text{ 为半径}, \quad i = 1, \cdots, n.$$

证 ☞ 令 x_i 为 A 的特征向量之一，大小为最大的那个分量. 从方程组 $(A-\lambda I)x=0$ 的第 i 个方程，我们有

$$(a_{ii}-\lambda)x_i = -\sum_{j\neq i} a_{ij}x_j,$$

$$|a_{ii}-\lambda| \leqslant \sum_{j\neq i} |a_{ij}| \left|\frac{x_j}{x_i}\right| \leqslant \sum_{j\neq i} |a_{ij}|,$$

它就是这个定理.

26.51 置换矩阵在每一行和每一列上只有一个 1，其他的都为零，关于它的特征值 Gerschgorin 定理告诉我们什么？

解 ☞ 这些圆或是以 0 为中心半径为 1，或是中心在 1 半径为零. 所以特征值位于离原点的一个单位范围内. 例如，

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

的特征值为 1 的立方根. 特别，恒等矩阵的特征值必定在中心为 1 半径为零的圆中.

26.52 Gerschgorin 定理对于具有优对角线的矩阵特别有用. 将它应用于矩阵

$$\begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & -1 & -1 \\ -1 & -1 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix}.$$

解 ☞ 所有的特征值必定落在一个中心在 4 半径为 3 的圆中. 由对称性，它们也必须为实的.

幂方法

26.53 什么是产生一个矩阵的最大特征值和特征向量的幂方法？

解 ☞ 假设矩阵 A 是 $n\times n$ 的具有 n 个独立向量 $V_1, V_2, \cdots, V_n$ 及一个真正的最大特征值 λ_1：$|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$. 则一个任意向量 V 可以表示为特征向量的组合，

$$V = a_1V_1 + a_2V_2 + \cdots + a_nV_n.$$

由此得

$$AV = a_1AV_1 + a_2AV_2 + \cdots + a_nAV_n = a_1\lambda_1V_1 + a_2\lambda_2V_2 + \cdots + a_n\lambda_nV_n,$$

连续乘以 A 我们到达

$$\begin{aligned} A^pV &= a_1\lambda_1^pV_1 + a_2\lambda_2^pV_2 + \cdots + a_n\lambda_n^pV_n \\ &= \lambda_1^p\left[a_1V_1 + a_2\left(\frac{\lambda_2}{\lambda_1}\right)^pV_2 + \cdots + a_n\left(\frac{\lambda_n}{\lambda_1}\right)^pV_n\right], \end{aligned}$$

假设 $a_1\neq 0$. 由于 λ_1 为最大的，方括号中的所有除第一项外的项均有零极限. 假如我们取任何 $A^{p+1}V$ 及 A^pV 对应分量的比，因此这个比应该有极限 λ_1. 此外，$\lambda_1^{-p}A^pV$ 将收敛于特征向量 a_1V_1.

26.54 对题 26.47 中使用过的矩阵应用幂法来寻找它的最大的特征值和特征向量.

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}.$$

解 ☞ 选择初始向量 $V=(1,1,1)$. 于是 $AV=(1,0,1)$ 及 $A^2V=(2,\ \ 2,2)$. 这里以 2 来除它是方便的，而在未来，为了保持数字合理我们继续以某些适当的因子去除. 用这个方法我们得到

$$A^7V = c(99, -140, 99), \qquad A^8V = c(338, -478, 338),$$

其中 c 是某个因子，分量的比为

$$\frac{338}{99}\approx 3.41414,\qquad \frac{478}{140}\approx 3.41429,$$

因而我们已经接近准确值 $\lambda_1=2+\sqrt{2}\approx 3.414214$. 我们最后的输出向量除以 338，它近似地变成 $(1,-1.41420,1)$，它接近在题 26.47 中得到的准确值 $(1,-\sqrt{2},1)$.

26.55　什么是 Rayleigh 商以及它可以怎样地使用于寻找主特征值?

解　☞ Rayleigh 商是 x^TAx/x^Tx，其中 T 表示转置. 假如 $Ax=\lambda x$，它就与 λ 相重. 假如 $Ax\approx\lambda x$ 那么可以想象 Rayleigh 商近似于 λ. 在一定的环境下，关于由幂方法产生的逐次向量的 Rayleigh 商收敛于 λ. 例如，令 x 为上题最终输出的向量，$(1,-1.41420,1)$. 这时

$$Ax=(3.41420,-4.82840,3.41420),$$

$$x^TAx=13.65672,\quad x^Tx=3.99996,$$

而 Rayleigh 商是近似地为 3.414214. 它准确到 6 位小数. 提示这里收敛到 λ_1 比分量比来得更快.

26.56　假设所有特征值都是实的，其他极特征值怎样才能找到?

解　☞ 若 $Ax=\lambda x$，则 $(A-qI)x=(\lambda-q)x$. 这意味着 $\lambda-q$ 是 $A-qI$ 的一个特征值. 通过适当地选择 q，比方说 $q=\lambda$，我们使其他极特征值占优并且幂方法可以被应用. 对于题 26.55 中的矩阵我们可以选 $q=4$ 并考虑

$$A-4I=\begin{bmatrix}-2 & -1 & 0\\ 1 & -2 & 1\\ 0 & -1 & -2\end{bmatrix},$$

还是取 $V=(1,1,1)$，我们立得关于向量 $(1,1.41421,1)$ 的 Rayleigh 商 -3.414214，这向量基本上是 $(A-4I)^8V$. 加上 4 我们有 0.585786，它是另一个极特征值 $2-\sqrt{2}$，准确到 6 位. 该向量也接近于准确的特征向量 $(1,\sqrt{2},1)$.

26.57　如何由幂法求绝对值最小的特征值?

解　☞ 若 $Ax=\lambda x$，则 $A^{-1}x=\lambda^{-1}x$. 这意味着 A 的绝对值最小的特征值可以作为 A^{-1} 的最大特征值 λ 的倒数求得. 关于题 26.55 的矩阵我们首先求得

$$A^{-1}=\frac{1}{4}\begin{bmatrix}3 & 2 & 1\\ 2 & 4 & 2\\ 1 & 2 & 3\end{bmatrix},$$

仍选取 $V=(1,1,1)$，但是现在使用 A^{-1} 代替 A，立得关于向量 $(1,1.41418,1)$ 的 Rayleigh 商 1.707107. 倒数商为 0.585786，所以我们仍有这个在题 26.47 及 26.56 中得到的特征值与特征向量. 寻找 A^{-1} 原本不是简单的任务，但是这个方法有时是对求绝对值最小之特征值的最好方法.

26.58　怎样才可以通过对出发向量 V 的一个适当选择来得到下一个最大特征值?

解　☞ 已经提出过各种算法，具有不同程度的成功. 困难在于把主特征值本身降到次要地位并保持它的次要地位. 舍入误差损害了若干理论上合理的方法，它使主特征值回到计算的主线上而遮掩了次主值，或者它限制了用于确定次主值的精度. 例如，假设在题 26.53 的讨论中，能安排出发向量 V 使 a_1 为零. 于是 λ_1 及 V_1 实际上永远不会出现，而且假若 λ_2 优于余下的那些特征值，它就担任以前由 λ_1 扮演的角色，因而相同的理由证明对 λ_2 和 V_2 的收敛. 用我们在题 26.54 中的矩阵，这一点可以很好地得到说明. 由于它是实的与对称的，故这矩阵就具有诸特征向量彼此正交的性质.（题 26.47 可以快捷地验证它.）这意味着 $V_1^TV=a_1V_1^TV_1$，因而假如 V 正交于 V_1，a_1 将为零. 假设我们取 $V=(-1,0,1)$，它正交 V_1，我们立刻得到 $AV=(-2,0,2)=2V$. 于是我们有精确的 $\lambda_2=2$ 及 $V_2=(-1,0,1)$. 然而，这里我们对出发向量的选择是碰巧的.

观察一下用一个合理但不是这么巧合的 V 会发生什么情况，几乎是件有趣的事. 比方说取 $V=(0,1,1.4142)$，它也像所要求的那样与 V_1 正交，接着我们发现 $A^3V\approx 4.8(-1,0.04,1.20)$，它多少有些类似于 V_2，并从它出发 Rayleigh 商提供令人满意的 $\lambda_2\approx 1.996$. 然而，在这之后计算恶化了而且最后我们得到 $A^{20}V\approx c(1,-1.419,1.007)$，它提供给我们的又一次是对 λ_1 及 V_1 的好的近似. 舍入误差使主特征值又活动起来. 不怕麻烦，稍稍改变每个向量 A^pV，使它正交于 V_1，可以造就一个较好的结果. 其他使用几个出发向量的办法也被尝试过的.

26.59 推出逆幂法.

解 ☞ 这是用于题 26.56 中的特征值移位的一个推广. 若 A 有特征值 λ_i, 则 $A-tI$ 及 $(A-tI)^{-1}$ 分别地有特征值 λ_i-t 与 $(\lambda_i-t)^{-1}$. 像在题 26.53 中那样应用幂法, 但是使用 $(A-tI)^{-1}$ 来代替 A, 我们有

$$(A-tI)^{-p}V=a_1(\lambda_1-t)^{-p}V_1+\cdots+a_n(\lambda_n-t)^{-p}V_n.$$

若 t 靠近一个特征值 λ_k, 并假设 $a_k\neq 0$ 且 λ_k 是一个孤立的特征值, 则 $a_k(\lambda_k-t)^{-p}V_k$ 这一项将在这个和中起决定作用. 要计算的幂就将引向一个 A 的特征值. 因为所有这些矩阵都有相同的特征向量. 这就是逆幂方法的基础.

这个思想的一个有趣的变化是使用值 t_i 的一个序列. 给出一个特征向量的初始近似, 比方说 $x^{(0)}$, 逐次计算

$$t_{i+1}=\frac{x^{(i)T}Ax^{(i)}}{x^{(i)T}x^{(i)}},\qquad x^{(i+1)}=c_{i+1}(A-t_{i+1}I)^{-1}x^{(i)},$$

t_{i+1} 是 Rayleigh 商, 作为 λ_k 的估计值, 而 $x^{(i+1)}$ 是对 V_k 的逼近. 收敛性曾在不同的假说下得到证明. 因子 c_{i+1} 选得使 $\|x^{(i+1)}\|=1$ 对某种范数成立.

实际上没必要去计算逆矩阵. 所需要的是由

$$w^{(i+1)}=(A-t_{i+1}I)^{-1}x^{(i)}$$

定义的向量 $w^{(i+1)}$, 所以, 通过解方程组

$$(A-t_{i+1}I)w^{(i+1)}=x^{(i)}$$

来得到它是更为经济的. 于是有 $x^{(i+1)}=c_{i+1}w^{(i+1)}$. 当序列发展时矩阵 $A-t_{i+1}I$ 将趋向奇异, 提示这方法也许有一个危险特征, 但是注意规格化及选主元, 能够得到精确结果.

26.60 什么是反迭代?

解 ☞ 给出一个对 A 的一个特征值的精确逼近, 反迭代是一个得到相应特征向量的快速方法. 令 t 为 λ 的一个近似值, 是从特征多项式或其他只产生特征值的方法得到的. 于是 $A-tI$ 接近奇异, 但像在题 26.8 中那样还是有一个因子分解:

$$P(A-tI)=LU,\qquad A-tI=P^{-1}LU.$$

正像在上题中那样, 我们以迭代公式

$$(A-tI)x^{(1)}=P^{-1}LUx^{(1)}=x^{(0)}$$

开始, 使用 $x^{(0)}$, 它在对应于 λ 之特征向量 x 的方向有一个非零分量. 选择 $x^{(0)}=P^{-1}L(1,1,\cdots,1)^T$ 有时为适当的, 或是同一回事

$$Ux^{(1)}=(1,1,\cdots,1)^T.$$

26.61 应用反迭代于题 26.47 的矩阵, 使用 0.586 作为对特征值 $2-\sqrt{2}$ 的一个近似. 由于特征向量 $x=(1,\sqrt{2},1)$ 已经获得, 这将充当对方法的潜能的一个小规模的例证.

解 ☞ 作为开始我们需要二个因子 L 及 U, 它们证明就是下面的

$$L=\begin{bmatrix}1&0&0\\-0.70721&1&0\\0&-1.4148&1\end{bmatrix},\qquad U=\begin{bmatrix}1.414&-1&0\\0&0.7068&-1\\0&0&-0.0008\end{bmatrix},$$

在这个例子中 $P=I$. $Ux^{(1)}=(1,1,\cdots,1)^T$ 的解, 由向后回代可以获得为 $x^{(1)}=(1250,1767,-1250)^T$. 接着

$$LUx^{(2)}=x^{(1)}$$

产生五位数字的 $x^{(2)}=(31,319,44,273,31,265)^T$. 规格化也就带来近似特征向量 $(1,1.414,0.998)^T$.

转化为正则形式

26.62 线代数的一个基本定理阐明: 一个实对称矩阵 A 只有实特征值并且存在着一个实正交矩阵 Q, 使 $Q^{-1}AQ$ 为一个对角阵. 于是对角元素就是特征值, 而 Q 的列就是特征向量. 推导产生这个正交矩阵 Q 的 Jacobi 公式.

解 ☞ 在 Jacobi 方法中 Q 是作为形如

$$Q_1 = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

的"旋转"矩阵的无穷乘积而得到的，Q_1 的所有其他元素与单位矩阵 I 的那些相同. 假如所表示的 4 个元素是在位置 $(i,i),(i,k),(k,i)$ 及 (k,k) 上的，那么相应的 $Q_1^{-1}AQ_1$ 的元素可以方便地计算为

$$b_{ii} = a_{ii}\cos^2\phi + 2a_{ik}\sin\phi\cos\phi + a_{kk}\sin^2\phi,$$

$$b_{ki} = b_{ik} = (a_{kk} - a_{ii})\sin\phi\cos\phi + a_{ik}(\cos^2\phi - \sin^2\phi),$$

$$b_{kk} = a_{ii}\sin^2\phi - 2a_{ik}\sin\phi\cos\phi + a_{kk}\cos^2\phi,$$

选择 ϕ 让 $\tan2\phi = 2a_{ik}/(a_{ii} - a_{kk})$，从而使得 $b_{ik} = b_{ki} = 0$. Jacobi 算法的每一步使一对非对角元素为零. 不幸的是下一步，当它制造出一对新的零时，在先前零的位置上引进了非零贡献. 尽管如此，逐次的形如 $Q_2^{-1}Q_1^{-1}AQ_1Q_2$ 的矩阵，等等，将趋向所要求的对角形式而且 $Q = Q_1Q_2\cdots$.

26.63 应用 Jacobi 方法于 $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$.

解 ☞ 取 $i=1, k=2$ 我们有 $\tan2\phi = -2/0$ 对它的解释是表示 $2\phi = \pi/2$. 于是 $\cos\phi = \sin\phi = 1/\sqrt{2}$ 以及

$$A_1 = Q_1^{-1}AQ_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 3 & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 2 \end{bmatrix}.$$

接下来我们取 $i=1, k=3$ 使 $\tan2\phi = -\sqrt{2}/(-1) = \sqrt{2}$，于是 $\sin\phi \approx 0.45969, \cos\phi \approx 0.88808$ 并且我们计算

$$A_2 = Q_2^{-1}A_1Q_2 = \begin{bmatrix} 0.88808 & 0 & 0.45969 \\ 0 & 1 & 0 \\ -0.45969 & 0 & 0.88808 \end{bmatrix} A_1 \begin{bmatrix} 0.88808 & 0 & -0.45969 \\ 0 & 1 & 0 \\ 0.45969 & 0 & 0.88808 \end{bmatrix}$$

$$= \begin{bmatrix} 0.63398 & -0.32505 & 0 \\ -0.32505 & 3 & -0.62797 \\ 0 & -0.62797 & 2.36603 \end{bmatrix}.$$

非对角线并未开始收敛于零，但是至少已开始减小. 在 9 次这类旋转后我们达到

$$A_9 = \begin{bmatrix} 0.58578 & 0.000000 & 0.000000 \\ .00000 & 2.00000 & 0.00000 \\ 0.00000 & 0.00000 & 3.41421 \end{bmatrix},$$

在这里早些时得到的特征值重又出现. 我们还有

$$Q \approx Q_1Q_2\cdots Q_9 = \begin{bmatrix} 0.50000 & 0.70710 & 0.50000 \\ 0.70710 & 0.00000 & -0.70710 \\ 0.50000 & -0.70710 & 0.50000 \end{bmatrix}.$$

在这里特征向量也是明显的.

26.64 关于一个实对称矩阵的 Jacobi 旋转算法的 Givens 的变更的三个主要部分是什么?

解 ☞ 在算法的第一部分中使用旋转将矩阵转化为三对角形式，只有主对角线和它的二条邻对角线不同于零. 第一次旋转是在(2,3)平面中，包含元素 a_{22}, a_{23}, a_{32} 及 a_{33}. 容易验证，这样一个旋转使用由 $\tan\phi = a_{13}/a_{12}$ 决定的 ϕ，元素 a_{13}(及 a_{31})换成了零，在$(2,i)$平面中的相继旋转将 a_{1i} 及 a_{i1} 换成 0，对 $i=4,\cdots,n$. ϕ 值由 $\tan\phi = a_{1i}/a'_{12}$ 所决定，其中 a'_{12} 表示当前位于 1 行，2 列上的元素. 接

下来是转变元素 $a_{24},\cdots,a_{2n}$，它们将被在(3,4)…(3,n)平面中的旋转变换成 0.以这种方式继续下去，一个三对角形式的矩阵将被得到，因为经我们操作所制造的零在以后的旋转中不会丢失.这可以由一个直接计算来证明并且使得 Givens 转化为有限过程，而 Jacobi 对角化是一个无限过程.

第二步包含形成序列

$$f_0(\lambda)=1,\quad f_i(\lambda)=(\lambda-\alpha_i)f_{i-1}(\lambda)-\beta_{i-1}^2f_{i-2}(\lambda),$$

其中诸 α 及 β 为我们新矩阵 B 的元素

$$B=\begin{bmatrix}\alpha_1&\beta_1&0&\cdots&0\\\beta_1&\alpha_2&\beta_2&\cdots&0\\0&\beta_2&\alpha_3&\cdots&0\\\cdots&\cdots&\cdots&\cdots&\beta_{n-1}\\0&0&0&\beta_{n-1}&\alpha_n\end{bmatrix}$$

与 $\beta_0=0$.这些 $f_i(\lambda)$ 证明为矩阵 $\lambda I-B$ 之主子矩阵的行列式，正如可以从

$$f_i(\lambda)=\begin{vmatrix}\lambda-\alpha_1&-\beta_1&0&\cdots&0\\-\beta_0&\lambda-\alpha_2&-\beta_2&\cdots&0\\0&-\beta_2&\lambda-\alpha_3&\cdots&0\\\cdots&\cdots&\cdots&\cdots&-\beta_{i-1}\\\cdots&\cdots&\cdots&-\beta_{i-1}&\lambda-\alpha_i\end{vmatrix}$$

看到的那样，通过沿着最后一列展开，有

$$f_i(\lambda)=(\lambda-\alpha_i)f_{i-1}(\lambda)+\beta_{i-1}D,$$

其中 D 在它的底行中只有元素 $-\beta_{i-1}$，因而等于 $D=-\beta_{i-1}f_{i-2}(\lambda)$.因此当 $i=n$ 我们在 $f_n(\lambda)$ 中有 B 的特征多项式.由于我们的旋转并不改变这个多项式，故这也是 A 的特征多项式.

现在，若某些 β_i 为零，行列式分裂为两个较小的行列式，它可以分开来处理.假如没有 β_i 为零，函数序列 $f_i(\lambda)$ 证明为 Sturm 序列(编号与题 25.53 给出的顺序相反).其结果是在一个给定的区间中特征值的个数可以用统计符号的变化次数来决定.

最后，第三步包含寻找特征向量.这里 B 的对角性质使 Gauss 消去法为直接地得到它的特征向量 U_j 的一个合理的过程(删去一个方程并赋予某个分量以任意值 1).于是 A 的相应的特征向量是 $V_j=QU_j$，其中 Q 再一次为我们旋转矩阵的乘积.

26.65 应用 Givens 方法于三阶的 Hilbert 矩阵

$$H=\begin{bmatrix}1&1/2&1/3\\1/2&1/3&1/4\\1/3&1/4&1/5\end{bmatrix}.$$

解 ☞ 对于这个小的矩阵只需一次旋转.取 $\tan\phi=2/3$，我们有 $\cos\phi=3/\sqrt{13}$ 及 $\sin\phi=2/\sqrt{13}$，于是

$$Q=\frac{1}{\sqrt{13}}\begin{bmatrix}\sqrt{13}&0&0\\0&3&-2\\0&2&3\end{bmatrix},$$

$$B=Q^{-1}HQ=\begin{bmatrix}1&\sqrt{13}/6&0\\\sqrt{13}/6&34/65&9/260\\0&9/260&2/195\end{bmatrix},$$

于是我们有了三对角矩阵.Sturm 序列由

$$f_0(\lambda)=1,\quad f_1(x)=\lambda-1,\quad f_2(\lambda)=\left(\lambda-\frac{34}{65}\right)(\lambda-1)-\frac{13}{16},$$

$$f_3(\lambda)=\left(\lambda-\frac{2}{195}\right)f_2(\lambda)-\frac{81}{67.600}(\lambda-1)$$

组成.它导出的 ± 号如表 26.3 所示.在 0 与 1 中有两个根及第三个根在 1 与 1.5 之间.迭代将这些位置更精化在 0.002688，0.122327，及 1.408319.特征值这样靠近零是该矩阵接近奇异的另外一个指标.

表 26.3

	f_0	f_1	f_2	f_3	变化
0	+	−	+	−	3
1	+	0	−	−	1
1.5	+	+	+	+	0

为寻找关于 λ_1 的特征向量,我们解 $BU_1=\lambda_1 U_1$ 很快就发现 $u_1=1, u_2=-1.6596, u_3=7.5906$ 是一种可能性.最后

$$V_1 = QU_1 = (1, -5.591, 5.395)^T,$$

它可以像所希望的那样被规格化.关于其他两个特征值的特征向量,相同的过程也适应于它们.

26.66 A 的一个相似变换定义为 $M^{-1}AM$,对任何非奇异矩阵 M.证明,这样的一种变换保持特征值不变.

证 ☞ 由于 $Ax=\lambda x$ 隐含了

$$MAM^{-1}(Mx) = \lambda(Mx)$$

我们立刻有 λ 是 MAM^{-1} 的一个特征值具有对应的特征向量 Mx.用于 Jacobi 及 Givens 方法中的正交变换是相似变换的特殊情况.

26.67 证明一个矩阵具有互不相同的特征值,与相应的独立特征向量时,可以通过一个相似变换将它转化为对角形式.

证 ☞ 通过使用 A 的特征向量作为列来构造矩阵 M,由此得

$$AM = MD,$$

其中 D 是对角阵并以特征值为它的对角元.因为特征向量均为线性独立的,M^{-1} 存在并且

$$M^{-1}AM = D$$

正如所求.将矩阵转化为特殊的,正则的形式的这个古典定理具有可疑的计算价值,因为寻找 M 看来要以整个问题的解作为先决条件.

26.68 什么是 Hessenberg 矩阵?

解 ☞ 它是这样一种矩阵,除了邻近主对角线的元素不为零外不是上三角为零就是下三角为零.假如上三角有零,该矩阵就是一个下 Hessenberg 阵,反之亦然.这里有二个小 Hessenberg 阵,第二个由于它的对称性还是一个三对角阵.

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

26.69 证明一个矩阵 A 可以通过 Guass 消去法加上一个相似变换转化为 Hessenberg 形式.

证 ☞ 假定我们以一个上 Hessenberg 阵作为我们的目标.所要求的下三角中的零可以逐列地在 $n-2$ 个步骤中生成.假设 $k-1$ 个这种步骤已经完成,并以 a'_{ij} 来表示新元素.然后 k 列中的零被安排如下:

(a)从元素 $a'_{k+1,k},\cdots,a'_{nk}$ 中找出绝对值最大的并将它所在行与 $k+1$ 行进行交换.这是部分主元步并且可以通过在当前的这个矩阵 A' 上左乘一个如题 26.8 中所介绍过的交换矩阵 $I_{r,k+1}$ 来完成.

(b)计算乘数

$$c_{jk} = -\frac{a''_{jk}}{a_{k+1,k}}, \quad j = k+2,\cdots,n$$

(带双撇号的代表交换后的元素).将 c_{jk} 乘 $k+1$ 行再与 j 行相加.这可以对所有的 j 同时来完成,通过将当前的这个 A'' 预乘以一个类似于题 26.8 中 L_i 的矩阵 G_k.

$$G_k=\begin{bmatrix}1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -c_{k+2,k} & 1 & \\ & & \vdots & & \ddots \\ & & -c_{nk} & & & 1\end{bmatrix} k+2\text{ 行},$$

k 列*

这是一个 Gauss 步.

(c)将当前的矩阵后乘 $I_{r,k+1}$ 及 G_k 的逆矩阵.这是相似变换的一步.当然,$I_{r,k+1}$ 就是它本身的逆,而 G_k 的逆可通过改变 c 元素的符号而得到.它完成了归纳法的第 k 个步骤,它可以概括为

$$G_kI_{r,k+1}A'I_{r,k+1}G_k^{-1},$$

以 A' 表示来自上一步的输入阵,或者当 $k=1$ 时就是 A 本身.

对 $k=1,\cdots,n-2$ 执行 a,b,c 三步,容易发现任何步上目标零被保留.

26.70 应用上题的算法于这个矩阵:

$$\begin{bmatrix}0 & 1 & 2 & 3\\ 2 & 3 & 0 & 1\\ 3 & 0 & 1 & 2\\ 1 & 2 & 3 & 0\end{bmatrix}.$$

解 ☞ 所有的要点呈现在图 26.3 中,这两个步骤左右排开.记住,作为一个预乘矩阵,$I_{r,k+1}$

$$I_{23}A\quad\begin{matrix}0 & 1 & 2 & 3\\ 3 & 0 & 1 & 2\\ 2 & 3 & 0 & 1\\ 1 & 2 & 3 & 0\end{matrix}\qquad\qquad I_{34}A'\quad\begin{matrix}0 & \frac{11}{3} & 1 & 3\\ 3 & \frac{5}{3} & 0 & 2\\ 0 & \frac{34}{9} & 2 & \frac{-2}{3}\\ 0 & \frac{11}{9} & 3 & \frac{-1}{3}\end{matrix}$$

$$G_1I_{23}A\quad\begin{matrix}0 & 1 & 2 & 3\\ 3 & 0 & 1 & 2\\ 0 & 3 & \frac{-2}{3} & \frac{-1}{3}\\ 0 & 2 & \frac{8}{3} & \frac{-2}{3}\end{matrix}\qquad\qquad G_2I_{34}A'\quad\begin{matrix}0 & \frac{11}{3} & 1 & 3\\ 3 & \frac{5}{3} & 0 & 2\\ 0 & \frac{34}{9} & 2 & \frac{-2}{3}\\ 0 & 0 & \frac{40}{17} & \frac{-2}{17}\end{matrix}$$

$$G_1I_{23}AI_{23}\quad\begin{matrix}0 & 2 & 1 & 3\\ 3 & 1 & 0 & 2\\ 0 & \frac{-2}{3} & 3 & \frac{-1}{3}\\ 0 & \frac{8}{3} & 2 & \frac{-2}{3}\end{matrix}\qquad\qquad G_2I_{34}A'I_{34}\quad\begin{matrix}0 & \frac{11}{3} & 3 & 1\\ 3 & \frac{5}{3} & 2 & 0\\ 0 & \frac{34}{9} & \frac{-2}{3} & 2\\ 0 & 0 & \frac{-2}{17} & \frac{40}{17}\end{matrix}$$

* 译注:原文有误.

$$G_1I_{23}AI_{23}G_1^{-1}\ (=A')\quad \begin{matrix} 0 & \frac{11}{3} & 1 & 3 \\ 3 & \frac{5}{3} & 0 & 2 \\ 0 & \frac{11}{9} & 3 & \frac{-1}{3} \\ 0 & \frac{34}{9} & 2 & \frac{-2}{3} \end{matrix} \qquad G_2I_{34}A'I_{34}G_2^{-1}\quad \begin{matrix} 3 & \frac{11}{3} & \frac{113}{34} & 1 \\ 3 & \frac{5}{3} & 2 & 0 \\ 0 & \frac{34}{9} & \frac{-1}{51} & 2 \\ 0 & 0 & \frac{186}{289} & \frac{40}{17} \end{matrix}$$

$$I_{23}\quad \begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{matrix} \qquad I_{34}\quad \begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{matrix}$$

$$G_1\quad \begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{-2}{3} & 1 & 0 \\ 0 & \frac{-1}{3} & 0 & 1 \end{matrix} \qquad G_2\quad \begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{-11}{34} & 1 \end{matrix}$$

图 26.3

交换行但是作为它自己的逆，而后乘它就交换列.给定的矩阵 A 是非对称的所以其结果是 Hessenberg 矩阵而不是三对角的.相似变换 MAM^{-1}的矩阵 M 是 $G_2I_{34}G_1I_{23}$.

26.71 什么是寻找特征值的 QR 方法?

解 ☞ 假设我们有一个上 Hessenberg 矩阵 H 并且可以将它分解为

$$H = QR,$$

具有 Q 为正交阵而 R 为上(或是右?)三角阵.在这个算法中归结为我们实际上首先要找的是

$$Q^TH = R,$$

通过逐次旋转将 H 转化为一个三角形式.定义

$$H^{(2)} = RQ = Q^THQ,$$

并注意到 $H^{(2)}$因为有题 26.66 中的定理将与 H 有相同的特征值.(由于 Q 是正交的，故 $Q^T = Q^{-1}$.)结果为 $H^{(2)}$也是 Hessenberg 阵，所以这个过程可以重复到由 $H^{(k)}$生成 $H^{(k+1)}$，以 H 作为 $H^{(1)}$以及 $k=1,\cdots$.收敛图是十分复杂的，但是在不同的假设下对角元趋于特征值同时下三角趋于零.(当然，R 因子在每个步骤上都是上三角，但在形成乘积 RQ 时，为重新恢复原来的特征值，次对角元素再一次变成非零.)这是 QR 方法的基本思想，下三角最后消失.

26.72 QR 方法第k 步上所要求的$Q^{(k)}$矩阵是怎样得到的? 即找 $Q^{(k)}$使

$$H^{(k+1)} = Q^{(k)T}H^{(k)}Q^{(k)}$$

对 $k=1,\cdots,$.

解 ☞ 完成它的一种方法是使用旋转，非常像在题 26.64 中提供的 Givens 方法.由于我们假设 H 为一个上 Hessenberg 阵，我们只需注意元素 $h_{i+1,i}$，对 $i=1,\cdots,n-1$，但是 $h_{i+1,i}$可以使用旋转

$$S_i^T = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & \cos\phi & \sin\phi & \\ & & -\sin\phi & \cos\phi & \\ & & & & \ddots \\ & & & & & 1 \end{bmatrix} \begin{matrix} \\ \\ i\text{ 行} \\ i+1\text{ 行} \\ \\ \\ \end{matrix}$$

（$\cos\phi$ 位于第 i 列，$\sin\phi$ 位于第 $i+1$ 列：i 列　$i+1$ 列）

将它变成零，并且计算 S_i^TH，假设 $\tan\phi = h_{i+1,i}/h_{i,i}$.(更为容易的是令 $\sin\phi = ch_{i+1,i}$，$\cos\phi = ch_{i,i}$并且选择c 使它的平方和为 1.)于是这些旋转的乘积

$$Q^T = S_{n-1}^T\cdots S_1^T$$

便是我们所要的.同样的论证可用于任何步骤上,所以上标(k)在这里被略掉了.

26.73 出现在题26.56中的特征值移位思想是如何应用于加速QR算法的收敛的?

解 ☞ 代替对矩阵H的因子分解,我们尝试对某个适当的p值转化

$$Q^T(H-pI)=R,$$

因此隐含着因子分解$H-pI=QR$.于是

$$Q^T(H-pI)Q=RQ=H^{(2)}-pI,$$

这展示了倒过来的乘积,它是方法的核心并且还定义了$H^{(2)}$.但是这样一来

$$H^{(2)}=Q^T(H-pI)Q+pI=Q^THQ.$$

所以$H^{(2)}$仍有与H相同的特征值.以手边的$H^{(2)}$,我们准备开始下一个迭代.p选得接近特征值是好的,然而在缺少这类内在信息的情况下推荐下面的选择.寻找当前的H的右下角的2×2子矩阵的特征值,并且当这些特征值是实的时,令P等于最接近h_{nn}的一个特征值.如果它们是复的,则令p为它们公共的实部.

26.74 给定一个小型的Hessenberg矩阵

$$H=\begin{bmatrix}4&2&1\\0&1&0\\0&2&3\end{bmatrix},$$

以QR方法寻找其特征值.

解 ☞ 容易发现它的特征值是对角元素4,1,3.但是观察QR方法怎样执行三角化还是有意义的.选择一个3的移位.我们计算

$$H-3I=\begin{bmatrix}1&2&1\\0&-2&0\\0&2&0\end{bmatrix}$$

它只需一次旋转就达到三角形式

$$S^T=\frac{1}{\sqrt{2}}\begin{bmatrix}\sqrt{2}&0&0\\0&-1&1\\0&-1&-1\end{bmatrix},\qquad S^T(H-3I)=\frac{1}{\sqrt{2}}\begin{bmatrix}\sqrt{2}&2\sqrt{2}&\sqrt{2}\\0&4&0\\0&0&0\end{bmatrix},$$

以S后乘就完成了相似变换

$$S^T(H-3I)S=\frac{1}{2}\begin{bmatrix}2&-\sqrt{2}&-3\sqrt{2}\\0&-4&-4\\0&0&0\end{bmatrix}.$$

最后我们加$3I$并得到

$$H^{(2)}=\begin{bmatrix}4&-\frac{\sqrt{2}}{2}&-\frac{3\sqrt{2}}{2}\\0&1&-2\\0&0&3\end{bmatrix},$$

保持了三角形式.一般来说,这种情况不会发生,需要好几个像上面这样的步骤.

26.75 应用QR方法于Hessenberg矩阵

$$H=\begin{bmatrix}4&1&1&1\\1&4&1&1\\0&1&4&1\\0&0&1&4\end{bmatrix},$$

对它来说精确特征值为6,4,3及3.

解 ☞ 大量的反复旋转最终将矩阵转化成下面的三角阵:

$$\begin{bmatrix}5.99997&1.50750&-0.17830&0.29457\\&3.99997&-0.44270&0.22152\\&&3.00098&-0.60302\\&&&2.99895\end{bmatrix},$$

在该矩阵中特征值显然就是沿着对角线的. 对于较大的作业当一个次对角线的元素为零时计算时间的节约将通过降阶来实现. 单纯观看下三角部分慢慢地消失是有趣的. 用上面的近似特征值直接得到对应的特征向量, 它们与准确值$(3,3,2,1)$, $(-1,-1,0,1)$及$(0,0,-1,1)$相匹配大约到3位小数上下. 没有第4个特征向量.

26.76 应用 QR 方法于三对角矩阵

$$\begin{bmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix},$$

然后用所得结果去"猜"准确特征值.

解 ☞ 为使具有该结果的进程运行起来再次允许反复地作旋转变换. 非对角元基本为零.

$$\begin{bmatrix} 5.618031 & & & \\ & 4.618065 & & \\ & & 3.381945 & \\ & & & 2.381942 \end{bmatrix}$$

由于给定的矩阵是对称的, 下三角与上三角二者均变成零. 十分明显留下特征值, 取最大的一个, 特征向量的直接计算得到

$$(1.00002, 1.61806, 1.61806, 1),$$

第4个分量是事先加以固定的. 猜想它应该是$(1, x, x, 1)$, 很快地导出方程

$$\lambda = x + 4, \qquad x^2 - x - 1 = 0.$$

这第二个方程由于它与 Fibonacci 数的联系而被人所熟悉. 根 $x=(1+\sqrt{5})/2$ 现在与 $\lambda=(9+\sqrt{5})/2$ 配对, 而 $x=(1-\sqrt{5})/2$ 与 $\lambda=(9-\sqrt{5})/2$ 配对, 给了我们精确解中的两个. 其他两个可类似地得到.

复方程组

26.77 解复方程组的问题怎样才能代之以解实方程组的问题?

解 ☞ 这几乎是自动实现的, 因为复数精确地相等的条件是它们的实部和虚部分别精确地相等. 方程

$$(A+iB)(x+iy) = a+ib$$

立刻等价于

$$Ax - By = a, \qquad Ay + Bx = b.$$

而这可以写成矩阵的形式

$$\begin{bmatrix} A & -B \\ B & A \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}.$$

一个复的 $n\times n$ 方程组为 $2n\times 2n$ 个实的系统来取代, 于是我们对实方程组的任何方法均可以被应用. 还可能以二个实的方程组来代替它

$$(B^{-1}A + A^{-1}B)x = B^{-1}a + A^{-1}b,$$
$$(B^{-1}A + A^{-1}B)y = B^{-1}b - A^{-1}a.$$

它们都是 $n\times n$ 的具有相同的系数矩阵. 这是从

$$(B^{-1}A + A^{-1}B)x = B^{-1}(Ax - By) + A^{-1}(Bx + Ay) = B^{-1}a + A^{-1}b,$$
$$(B^{-1}A + A^{-1}B)y = B^{-1}(Ay + Bx) + A^{-1}(By - Ax) = B^{-1}b - A^{-1}a$$

得到的. 使用这些较小的方程组略微缩短了整体的计算.

26.78 将复矩阵的求逆简化为实矩阵的求逆.

解 ☞ 令给定的矩阵为 $A+iB$ 以及它的逆矩阵为 $C+iD$. 我们要寻求 C 及 D 使得$(A+iB)(C+iD)=I$ 成立. 假定 A 是非奇异的于是 A^{-1}存在. 此时

$$C = (A + BA^{-1}B)^{-1}, \qquad D = -A^{-1}B(A + BA^{-1}B)^{-1}.$$

这可以通过直接回代得到验证. 如果 B 是非奇异的, 则

$$C = B^{-1}A(AB^{-1}A + B)^{-1}, \qquad D = -(AB^{-1}A + B)^{-1}.$$

这也可以通过回代得到验证. 假如 A 与 B 都是非奇异, 这二个结果当然是恒等的. 在 A 与 B 都是奇异的情况下, 而 $(A+iB)$ 不是奇异的, 此时一个更为复杂的过程看来是必要的. 首先决定一个实数 t 使得矩阵 $E=A+tB$ 为非奇异的. 然后, 取 $F=B-tA$, 我们得到 $E+iF=(1-it)(A+iB)$ 于是

$$(A+iB)^{-1}=(1-it)(E+iF)^{-1}.$$

由于 E 是非奇异的故可以用第一种方法加以计算.

26.79 把寻找特征值和特征向量的 Jacobi 方法拓广到 Hermite 矩阵的情况.

解 我们利用 Hermitian 矩阵在一个酉变换下成为对角化的事实, 即 $U^{-1}HU$ 是一个对角阵. 矩阵 H 及 U 有 $\overline{H}^T=H$ 及 $\overline{U}^T=U^{-1}$ 的性质. 矩阵 U 可以通过形如

$$U_1=\begin{bmatrix}\cos\phi & -\sin\phi e^{-i\theta}\\ \sin\phi e^{i\theta} & \cos\phi\end{bmatrix}$$

所有其他元素与 I 的那些一致的矩阵的无穷乘积而得到. 所示的 4 个元素是在 $(i,i),(i,k),(k,i)$ 及 (k,k) 的位置上. 假如 H 的相应的元素为

$$H=\begin{bmatrix}a & b-ic\\ b+ic & d\end{bmatrix},$$

将 ϕ 及 θ 选得使

$$\tan\theta=\frac{c}{b},\qquad \tan2\phi=\frac{2(b\cos\theta+c\sin\theta)}{a-d},$$

则 $U^{-1}HU$ 的 (i,k) 及 (k,i) 元素会有实部及虚部都为零,

$$(d-a)\cos\phi\sin\phi\cos\theta+b\cos^2\phi-b\sin^2\phi\cos2\theta-c\sin^2\phi\sin2\theta=0,$$

$$(a-d)\cos\phi\sin\phi\sin\theta-c\cos^2\phi+b\sin^2\phi\sin2\theta-c\sin^2\phi\cos2\theta=0.$$

这种类型的旋转像在题 26.62 中那样迭代地被使用直到所有非对角元素都已变得足够小, 这(实)特征值于是为结果的对角元素所逼近, 而特征向量则被 $U=U_1U_2U_3\cdots$ 的列所逼近.

26.80 对一个一般的复矩阵怎样才能找到它的特征值和特征向量? 假设所有特征值都是互异的.

解 作为第一步我们得到一个酉矩阵 U 使得 $U^{-1}AU=T$, 其中 T 是一个上三角阵, 主对角线下的所有元素为零. 这个 U 再次是由在上题中所示的形式 U_1 的旋转矩阵的无穷乘积而得到的, 现在我们将它写成

$$U_1=\begin{bmatrix}x & -\bar{y}\\ y & x\end{bmatrix},$$

于是在 $U_1^{-1}AU_1$ 的 (k,i) 位置上的元素为

$$a_{ki}x^2+(a_{kk}-a_{ii})xy-a_{ik}y^2,$$

为了使它为零我们令 $y=Cx$, $x=1/\sqrt{1+|C|^2}$. 它自动地向我们保证 U_1 将为酉的. 然后由条件 $a_{ik}C^2+(a_{ii}-a_{kk})C-a_{ki}=0$ 确定 C, 它使

$$C=\frac{1}{2a_{ik}}\left[(a_{kk}-a_{ii})\pm\sqrt{(a_{kk}-a_{ii})^2+4a_{ik}a_{ki}}\right],$$

两种符号均可使用, 更可取的是使 $|C|$ 更小一些的那个符号. 这类旋转持续地做下去直到在主对角线下的元素基本上为零. 结果矩阵为

$$T=U^{-1}AU=\begin{bmatrix}t_{11} & t_{12} & \cdots & t_{1n}\\ 0 & t_{22} & \cdots & t_{2n}\\ \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & \cdots & t_{nn}\end{bmatrix},$$

其中 $U=U_1U_2\cdots U_N$. T 和 A 两者的特征值均为对角元素 t_{ii}.

下一步得到 T 的特征向量, 作为 W 的列

$$W=\begin{bmatrix}1 & w_{12} & w_{13} & \cdots & w_{1n}\\ 0 & 1 & w_{23} & \cdots & w_{2n}\\ 0 & 0 & 1 & \cdots & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & 0 & \cdots & w_{nn}\end{bmatrix},$$

第一列已经是属于 t_{11} 的特征向量. 为了使第二列为一个属于 t_{22} 的特征向量, 我们要求 $t_{11}w_{12}+t_{12}=t_{22}w_{12}$ 或是 $w_{12}=t_{12}/(t_{22}-t_{11})$, 假设 $t_{11}\neq t_{22}$. 类似地, 使第三列为一个特征向量我们要求

$$w_{23}=\frac{t_{23}}{t_{33}-t_{22}},\qquad w_{13}=\frac{t_{12}w_{23}+t_{13}}{t_{33}-t_{11}}.$$

一般来说 w_{ik} 由递推公式

$$w_{ik}=\sum_{j=i+1}^{k}\frac{t_{ij}w_{jk}}{t_{kk}-t_{ii}}$$

得到, 逐次地取 $i=k-1,k-2,\cdots,1$. 最后 A 自己的特征向量可由 UW 的列得到.

补　充　题

26.81　应用 Gauss 消去法来求这个方程组的解向量:

$$\begin{aligned}w+2x-12y+8z&=27,\\5w+4x+7y-2z&=4,\\-3w+7x+9y+5z&=11,\\6w-12x-8y+3z&=49.\end{aligned}$$

26.82　应用题 26.10 的方法来求这个方程组的解向量:

$$\begin{aligned}33x_1+16x_2+72x_3&=359,\\-24x_1-10x_2-57x_3&=281,\\-8x_1-4x_2-17x_3&=85.\end{aligned}$$

26.83　假设已经找到方程组

$$\begin{aligned}1.7x_1+2.3x_2-1.5x_3&=2.35,\\1.1x_1+1.6x_2-1.9x_3&=-0.94,\\2.7x_1-2.2x_2+1.5x_3&=2.70.\end{aligned}$$

有一个解靠近(1,2,3). 应用题 26.28 的方程来得到一个改进的近似值.

26.84　应用 Gauss 消去法于如下的方程组, 以有理形式进行计算, 因而没有舍入误差的引入并由此得到一个精确解. 系数矩阵是四阶的 Hilbert 阵.

$$\begin{aligned}x_1+\frac{1}{2}x_2+\frac{1}{3}x_3+\frac{1}{4}x_4&=1,\\\frac{1}{2}x_1+\frac{1}{3}x_2+\frac{1}{4}x_3+\frac{1}{5}x_4&=0,\\\frac{1}{3}x_1+\frac{1}{4}x_2+\frac{1}{5}x_3+\frac{1}{6}x_4&=0,\\\frac{1}{4}x_1+\frac{1}{5}x_2+\frac{1}{6}x_3+\frac{1}{7}x_4&=0.\end{aligned}$$

26.85　重复上题, 将所有系数均换成取 3 位有效数字的小数. 在整个计算过程中只保留 3 位有效数字. 试问你的结果与上题中的精确解接近程度如何? (高阶的 Hilbert 矩阵特别麻烦, 即使有许多位小数可以用来进行计算.)

26.86　应用 Gauss-Seidel 迭代于下面的方程组

$$\begin{aligned}-2x_1+x_2\qquad\qquad&=-1,\\x_1-2x_2+x_3\qquad&=0,\\x_2-2x_3+x_4&=0,\\x_3-2x_4&=0.\end{aligned}$$

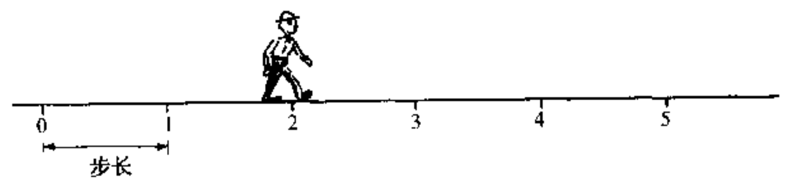


图 26.4

对所有 k 以逼近值 $x_k=0$ 开始,将方程组改写成每个方程就其主对角的未知量解出.在作了若干次迭代后你能猜出准确的解向量吗?这个问题可以用一个随机步行者来解释,他沿着直线走的每一步朝左或朝右是随机的如图 26.4 所示.当他到达一个端点时他便停下,每一个 x_k 值代表他从 k 位置上到达左端点的概率.我们可以定义 $x_0=1$ 及 $x_5=0$,在这种情况下每个方程有形式 $x_{k-1}-2x_k+x_{k+1}=0, k=1,\cdots,4$.

26.87　超松弛法是否朝着题 26.86 的精确解方向加速收敛?

26.88　应用 Gauss-Seidel 方法于方程组

$$x_k=\frac{3}{4}x_{k-1}+\frac{1}{4}x_{k+1},\qquad k=1,\cdots,19,$$

$$x_0=1,\qquad x_{20}=0.$$

它可以解释为表示一个随机步行者在一条有编号为 0 到 20 之位置的直线上,他朝左移动经常 3 倍于朝右.

26.89　上题是一个关于差分方程的边值问题,证明它的精确解为 $x_k=1-(3^k-1)/(3^{20}-1)$.对 $k=0(1)20$ 计算这些值并与以迭代算法所得结果进行比较.

26.90　对同样的方程组应用超松弛法.以 w 值作试验.低松弛法($w<1$)看起来是否适用于这个问题?

26.91　应用我们的任何一个方法于下面的方程组:

$$\begin{aligned}
x_1+x_2+x_3+x_4+x_5&=1,\\
x_1+2x_2+3x_3+4x_4+5x_5&=0,\\
x_1+3x_2+6x_3+10x_4+15x_5&=0,\\
x_1+4x_2+10x_3+20x_4+35x_5&=0,\\
x_1+5x_2+15x_3+35x_4+70x_5&=0.
\end{aligned}$$

26.92　以题 26.38 的消去算法求题 26.81 的系数矩阵的逆矩阵.

26.93　以交换法求同一矩阵的逆矩阵.

26.94　以任何一种我们的方法求题 26.86 中系数矩阵的逆矩阵.

26.95　尝试使用 3 位算术求 4 阶 Hilbert 矩阵的逆矩阵.

26.96　尝试对 Wilson 矩阵求逆,再对逆矩阵求逆,它能接近原始矩阵到什么程度?

$$\begin{bmatrix}10&7&8&7\\7&5&6&5\\8&6&10&9\\7&5&9&10\end{bmatrix}.$$

26.97　应用题 26.43 的方法于题 26.82 的矩阵.看起来它是否朝着精确逆矩阵收敛?

$$A^{-1}=\frac{1}{6}\begin{bmatrix}-58&-16&-192\\48&15&153\\16&4&54\end{bmatrix}.$$

26.98　求题 26.81 中系数矩阵的行列式的值.

26.99　求题 26.82 中系数矩阵的行列式的值.

26.100　什么是 4 阶 Hilbert 矩阵的行列式?

26.101　应用题 26.48 的方法求 $Ax=\lambda x$ 的特征值和特征向量,其中 A 是 3 阶 Hilbert 矩阵.使用有理算术并求精确的特征多项式.

26.102　参照题 26.101,应用同样的方法于

$$\begin{aligned}
(2-\lambda)x-x_2&=0,\\
-x_1+(2-\lambda)x_2-x_3&=0,\\
-x_2+(2-\lambda)x_3-x_4&=0,\\
-x_3+(2-\lambda)x_4-x_5&=0,\\
-x_4+(2-\lambda)x_5&=0.
\end{aligned}$$

26.103　使用幂法找矩阵

$$A=\begin{bmatrix}2&-1&0&0\\-1&2&-1&0\\0&-1&2&-1\\0&0&-1&2\end{bmatrix}$$

的最大特征值与特征向量

26.104　使用幂法求 3 阶 Hilbert 矩阵的最大特征值与特征向量.

26.105　应用 Jacobi 方法于 3 阶 Hilbert 矩阵.

26.106　应用 Jacobi 方法于题 26.103 的矩阵.

26.107　应用 Givens 方法于题 26.103.

26.108　应用 Givens 方法于 4 阶 Hilbert 矩阵.

26.109　解方程组

$$\begin{aligned} x_1 + ix_2 \qquad\quad &= 1, \\ -ix_1 + x_2 \quad + ix_3 &= 0, \\ -ix_2 \quad + x_3 &= 0. \end{aligned}$$

用题 26.77 的方法.

26.110　应用题 26.78 的方法求题 26.109 中系数矩阵的逆矩阵.

26.111　应用题 26.79 中提出来的 Jacobi 方法,求题 26.109 中系数矩阵的特征值与特征向量.

26.112　应用题 26.80 的算法于矩阵

$$A = \begin{bmatrix} 1 & i & -1 \\ i & 1 & i \\ -1 & i & 1 \end{bmatrix}.$$

26.113　假设矩阵 A 有一个 LU 因子分解,我们有题 26.14 中关于决定因子元素的公式.

$$u_{rj} = a_{rj} - l_{r1}u_{1j} - l_{r2}u_{2j} - \cdots - l_{r,r-1}u_{r-1,j}, \qquad j \geqslant r,$$
$$u_{rr}l_{ir} = a_{ir} - l_{i1}u_{1r} - l_{r2}u_{2r} - \cdots - l_{i,r-1}u_{r-1,r}, \qquad i > r,$$

假定它们都是从左到右计算的.以一撇表示遭受舍入误差的计算值,于是,u'_{rj}的计算像下面这样开始.(见题 1.22 及 1.23.)

$$a_{rj}(1+E) - l'_{r1}u'_{1j}(1+E)^{(2)},$$

每个 E 代表一个舍入误差,在不同的地方很可能有不同的误差,而上标不表示幂而是不同的$(1+E)$因子的编号,这种措施将本来要长的表达式缩短一些.继续下去,

$$a'_{rj}(1+E)^{(2)} - l'_{r1}u'_{1j}(1+E)^{(3)} - l'_{r2}u'_{2j}(1+E)^{(2)},$$

直到最终我们得到计算的 u'_{rj}:

$$u'_{rj} = a_{rj}(1+E)^{(r-1)} - l'_{r1}u'_{1j}(1+E)^{(r)} - \cdots - l'_{r,r-1}u'_{r-1,j}(1+E)^{(2)}.$$

证明相应的关于计算 l'_{ir}的表达式如下:

$$\begin{aligned} u'_{rr}l'_{ir}(1+E) = a_{ir}(1+E)^{(r-1)} - l'_{i1}u'_{1r}(1+E)^{(r)} \\ - \cdots - l'_{i,r-1}u'_{r-1,r}(1+E)^{(2)}. \end{aligned}$$

26.114　由下式定义 Δ_2

$$(1+E_1)(1+E_2) = 1 + 2\Delta_2,$$

并注意

$$|\Delta_2| = \left|\frac{1}{2}(E_1 + E_2 + E_1E_2)\right| \leqslant u + \frac{1}{2}u^2,$$

以 u 表最大舍入误差.类似地证明由 $(1+E_1)(1+E_2)(1+E_3) = 1 + 3\Delta_3$ 定义的 Δ_3 的界 $u + u^2 + \frac{1}{3}u^3$存在,以及更为一般地我们可以记

$$(1+E)^{(n)} = 1 + n\Delta_n.$$

Δ_n 以$[(1+u)^n - 1]/n$ 为界.

26.115　综合前两题的结果来得到(以 Δ 表一个适当的 Δ_k)

$$\begin{aligned} l'_{rk}u'_{kj} - a_{rj} &= a_{rj}(j-1)\Delta - l_{r1}u_{1j}j\Delta - \cdots - l_{rj}u_{jj}\Delta, \qquad r > j, \\ &= a_{rj}(r-1)\Delta - l_{r1}u_{1j}r\Delta - \cdots - l_{rr}u_{rj}\Delta, \qquad r \leqslant j. \end{aligned}$$

并注意到以矩阵形式它等价于

$$L'U' = A + F$$

具有 F 的元素如上述等式右侧所示.这说明因子分解 $L'U'$对被扰动的矩阵$A+F$是精确的.

26.116　证明上题中 F 的元素绝对值不超过$n\Delta$乘 A 与$L'U'$的组合项.也就是说

$$|f_{ij}| \leqslant n\Delta(|a_{ij}| + b_{ij}),$$

其中 Δ 为被包含的所有 Δ_k 的界，而 b_{ij} 是 L' 的 i 行元素的绝对值与 U' 的 j 列元素绝对值计算而来. 内部舍入误差效应的估计强烈地依赖于 b_{ij}. 这些可以在因子分解后再计算. 此处 n 为原始矩阵 A 的阶. 假设 n 不太大加上 b_{ij} 是配合的，我们由此推断全局误差为最大舍入误差的适中的倍数.

26.117 在题 26.9 中导出的向前与向后回代公式

$$y_r = b_r - l_{r1}y_1 - \cdots - l_{r,r-1}y_{r-1},$$

$$u_{ii}x_i = y_i - u_{i,i+1}x_{i+1} - \cdots - u_{in}x_n$$

具有与对舍入误差传播刚分析过的那些相同的形式. 理由与上题的颇为相同，人们可以为计算出 y' 得到方程

$$(L' + G)y' = b,$$

其中 $|g_{ij}| \leqslant n\Delta|l'_{ij}|$，然后为计算出解本身有

$$(U' + H)x' = y',$$

此处 $|h_{ij}| \leqslant n\Delta|u_{ij}|$.

将这些结果与上题的综合起来，证明它为

$$(A + E)x' = b,$$

其中 E 为 F, G, H, L 及 U 的混合体. 进一步推出估计式

$$|e_{ij}| \leqslant n\Delta[|a_{ij}| + (3 + n\Delta)|b_{ij}|],$$

其中 b_{ij} 定义如前.

26.118 应用题 26.80 的算法于实的非对称矩阵

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 5 \\ 1 & 4 & 7 \end{bmatrix}.$$

26.119 解方程组

$$\begin{aligned} 6.4375x_1 + 2.1849x_2 - 3.7474x_3 + 1.8822x_4 &= 4.6351, \\ 2.1356x_1 + 5.2101x_2 + 1.5220x_3 - 1.1234x_4 &= 5.2131, \\ -3.7362x_1 + 1.4998x_2 + 7.6421x_3 + 1.2324x_4 &= 5.8665, \\ 1.8666x_1 - 1.1104x_2 + 1.2460x_3 + 8.3312x_4 &= 4.1322. \end{aligned}$$

26.120 求下面方程组的所有特征值：

$$\begin{aligned} 4x + 2y + z &= \lambda x, \\ 2x + 4y + 2z &= \lambda y, \\ x + 2y + 4z &= \lambda z. \end{aligned}$$

26.121 求方程组

$$\begin{bmatrix} 4 & 2 & 2 \\ 2 & 5 & 1 \\ 2 & 1 & 6 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

的所有特征值与特征向量.

26.122 对 Pascal 矩阵求逆：

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 3 & 6 & 10 & 15 \\ 1 & 4 & 10 & 20 & 35 \\ 1 & 5 & 15 & 35 & 70 \end{bmatrix}.$$

26.123 对下面的矩阵求逆：

$$\begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{5} & \frac{1}{7} \\ \frac{1}{5} & \frac{1}{7} & \frac{1}{9} \end{bmatrix}.$$

26.124 对下面的矩阵求逆：

$$\begin{bmatrix} 5+i & 4+2i \\ 10+3i & 8+6i \end{bmatrix}.$$

26.125 求

$$\begin{bmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & 3 & 2 \end{bmatrix}$$

的最大特征值到三位.

26.126 求

$$\begin{bmatrix} 8 & -5i & 3-2i \\ 5i & 3 & 0 \\ 3+2i & 0 & 2 \end{bmatrix}$$

的最大特征值及相应的特征向量.

26.127 求

$$\begin{bmatrix} 9 & 10 & 8 \\ 10 & 5 & -1 \\ 8 & 6 & 3 \end{bmatrix}$$

的两个极端特征值.

26.128 证明矩阵

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

的特征多项式为 $\lambda^2-\lambda-1$ 并注意它与题 18.23 中以及在别处出现的 Fibonacci 数的关系.对于更为一般的 n 阶"Fibonacci"矩阵的特征多项式是什么？

用我们方法的任何一种求

$$F_n = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

的特征值.给出某个初始向量 x，当 $p=2,\cdots$时向量 $F_n^p x$ 是什么？

26.129 应用 QR 方法于 Hessenberg 矩阵：

$$\begin{bmatrix} 2 & 1 & 0.5 & 0.1 \\ 1 & 3 & 1 & 0.5 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 0.5 & 1 \end{bmatrix}.$$

26.130 应用 QR 方法于三对角矩阵：

$$\begin{bmatrix} 2.5 & -2.0 & 0 & 0 \\ -2.0 & 3.5 & 1.5 & 0 \\ 0 & 1.5 & 2.5 & -1.0 \\ 0 & 0 & -1.0 & 1.5 \end{bmatrix}.$$

26.131 将一个正方形顺时针方向旋转四分之一圈可以通过应用置换矩阵 R 于向量$(1,2,3,4)^T$ 来模拟.(见图 26.5.)相对于垂直线(虚线)的反射可以用矩阵 V 来模拟.易获 R 的特征值为 $1, i, -1, -i$，而 V 的特征值为 $1,1,-1,\ \ 1$.二个矩阵均是 Hessenberg 型的.题 26.73 的 QR 算法在两种情况中是否都将收敛？

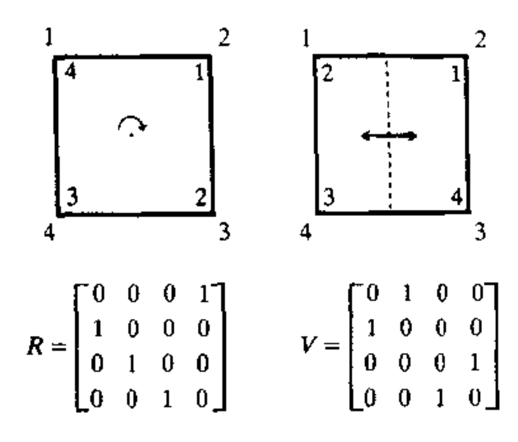

1 2
4 1
3 2
4 3
1 2
2 1
3 4
4 3
$R = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$
$V = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$

图 26.5

第二十七章　线性规划

基本问题

一个线性规划问题要求一个线性函数

$$H = c_1x_1 + \cdots + c_nx_n$$

在受到形如

$$a_{i1}x_1 + \cdots + a_{in}x_n \leqslant b_i \quad 0 \leqslant x_j$$

约束下极小化(或极大化)，其中 $i=1,\cdots,m$ 及 $j=1,\cdots,n$. 这个问题以向量形式可以写成

$$H(x) = c^{\mathrm{T}}x = \text{极小}, \qquad Ax \leqslant b, \quad 0 \leqslant x.$$

线性规划的一个重要定理说明所要求的极小值(或极大值)出现在一个极端可行点处. 一个点$(x_1,\cdots,x_n)$称为可行的假如它的坐标满足所有 $n+m$ 个约束，而一个极端可行点是这样的一个点，其中至少有 n 个约束真正成为等式. 松弛变量 $x_{n+1},\cdots,x_{n+m}$的引进将约束转化为

$$a_{i1}x_1 + a_{i2}x_2 - \cdots + a_{in}x_n + x_{n+i} = b_i$$

的形式，对 $i=1,\cdots,m$. 它允许一个极端可行点恒等于这样的一个点在该点上 n 个或更多的变量(包括松弛变量)为零. 这是一个重大的方便. 在特殊情况下多于一个的极端可行点可以提供所要求的极小，在这种情况下其他的可行点也对这个目的适用. H 的一个极小点称为解点.

单纯形方法是一个算法，对于在某些极端可行点出发，然后通过一系列交换，有规则地进行到其他的这种点直到获得一个解点. 它是以这样的一个方法来完成的，它稳定地使 H 的值减少. 所包含的交换过程基本上与前章关于矩阵求逆所列出的相同.

对偶定理是在两个问题解之间的关系

$$c^Tx = \text{极小} \qquad Ax \geqslant b, \quad 0 \leqslant x.$$

$$y^Tb = \text{极大} \quad y^TA \leqslant c^T, \quad 0 \leqslant y.$$

它们被称作对偶问题，而且它们包含了相同的数 a_{ij}, b_i, 及 c_j. 对应的极小值及极大值证明是相同的，不管单纯形方法应用在哪一个问题上(假定对二者中容易的一个). 允许两个问题的解都能在这些结果中引出，这显然是一个重大的方便.

两个相关的问题

1. **两人游戏**要求从下面的“支付”(“pay off”)矩阵

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

   中 R 选一个行而 C 选一个列. 元素 a_{ij} 位于所选的行与列的交叉处，它决定 R 必须付给 C 的量，自然 C 希望将他所期望赢得的极大化，与此同时 R 希望将它期望的损失极小化，这些矛盾观点导出一个对偶线性规划. 它可以用单纯形方法来解，这个解称为对两个游戏者的最佳策略.

2. **超定的线性方程组**. 在这个方程组中方程多于未知量，没有一个向量可以满足整个方程组，它可以作为线性规划问题来处理，在该问题中我们寻找向量 x 它在某种意义下有最小误差，其细节公布在第 28 章中.

题 解

单纯形方法

27.1 求满足不等式

$$0 \leqslant x_1, \quad 0 \leqslant x_2, \quad -x_1 + 2x_2 \leqslant 2, \quad x_1 + x_2 \leqslant 4, \quad x_1 \leqslant 3$$

的 x_1 及 x_2 并且使函数 $F = x_2 - x_1$ 极大化.

解 ☞ 由于只包含了两个变量,为方便起见整个问题可用几何来解释.在 x_1, x_2 的一个平面中 5 个不等式将点 (x_1, x_2) 约束到落入图 27.1 的阴影部分.在每一种情况下等号对应于 (x_1, x_2) 它在 5 条直的边界线段的一个上.在这些约束的限制下,将 F 极大化等价于找斜率为 1 有最大的 y 截距而且还与阴影区域相交的直线,看来很清楚所要求的直线 L_1 是 $1 = x_2 - x_1$ 而交点是(0,1),因此,对一个最大值来说,$x_1 = 0, x_2 = 1, F = 1$.

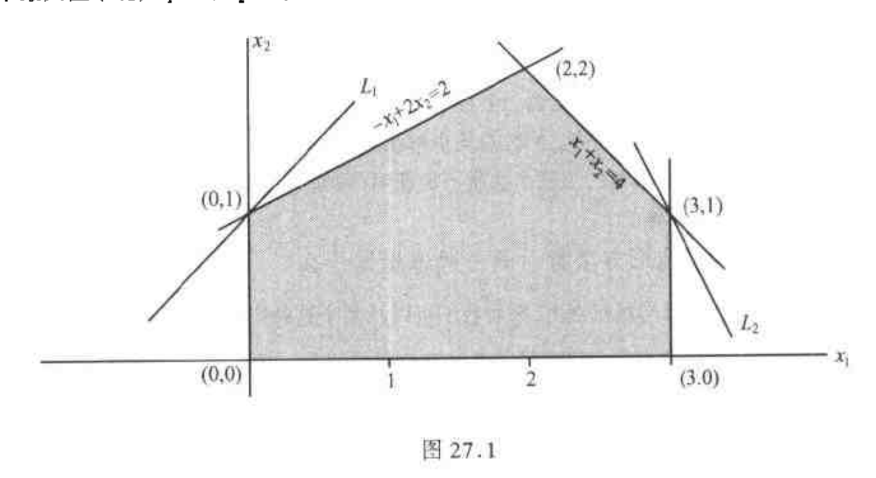


图 27.1

27.2 寻找满足与题 27.1 相同的不等式约束且使 $G = 2x_1 + x_2$ 为极大的 (x_1, x_2).

解 ☞ 现在我们找一条斜率为 −2 的直线,并具有最大的 y 截距同时与阴影部分相交,这根直线 L_2 是 $7 = 2x_1 + x_2$ 而所要求的点为 $x_1 = 3, x_2 = 1$.(参看图 27.1)

27.3 寻找 y_1, y_2, y_3,满足约束

$$0 \leqslant y_1, \quad 0 \leqslant y_2, \quad 0 \leqslant y_3, \quad y_1 - y_2 - y_3 \leqslant 1, \quad -2y_1 - y_2 \leqslant -1,$$

并使 $H = 2y_1 + 4y_2 + 3y_3$ 极小化.

解 ☞ 将整个问题用几何加以解释,我们发现 5 个不等式约束 (y_1, y_2, y_3) 点落入一个如图 27.2

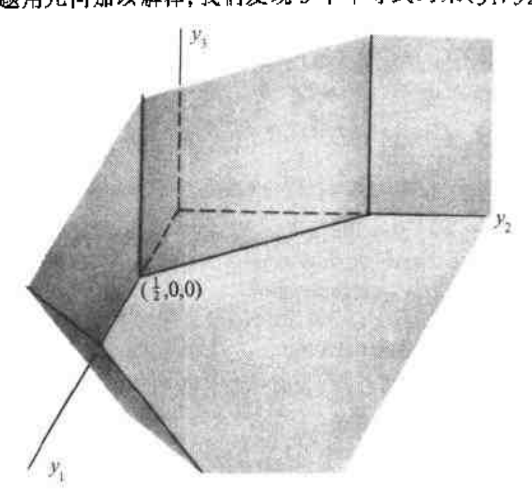


图 27.2

所示的区域,这个区域在 y_1,y_2,y_3 的正方向是无界的,然而,其他地方为 5 个平面所限的部分,如阴影所示,这些平面对应于在我们的 5 个约束中等号成立.在这些约束的限制下,极小化 H 等价于寻找一个平面具有法向量(2,4,3),有最小的截面但仍与所给的区域相交.易见这个平面就是 $1=2y_1+4y_2+3y_3$ 而交点为 $\left(\frac{1}{2},0,0\right)$.

27.4 列出线性规划问题的三个主要性质以及它们的解,它们为前面的问题所说明.

解 ☞ 令这个问题是寻找一个点 x 具有坐标 $(x_1,x_2,\cdots,x_n)$,受到约束 $0\leqslant x$, $Ax\leqslant b$ 而且将函数 $H(x)=c^{\mathrm T}_x=\sum c_ix_i$ 极小化.称一个满足所有这些约束的点是一个可行点(假如有任何这种点存在的话),则:

1. 可行点的集合是凸的,也就是说连接两个可行点的线段全部由可行点组成,这是由于每个约束都定义了一个半空间,而可行点的集合就是这些半空间的交集.
2. 存在某些极端可行点,凸集的角点,它为在这些点上至少有 n 个约束成为等式这个事实所确认.以二维的情况为例,精确地有 $n=2$ 个边界段在这类角点上相遇.以三维的情况为例,精确地有三个边界平面在每个这种角点上相遇.然而当 $n\geqslant3$ 时可能有更多的平面(或超平面)在一个角点上汇合到一起.
3. 解点常常是一个极端可行点.这是由于要极小化的函数 $H(x)$ 的线性.(有可能二个极端可行点都是解,在这种情况下整个的连结它们的边均由解组成,等等.)

在这里我们对线性规划的这三个性质不加证明.假如 $H(x)$ 是要被极大化的,或是约束读作 $Ax\geqslant b$,它们同样是正确的.

27.5 关于解线性规划的单纯形方法背后的一般思想是什么?

解 ☞ 由于解出现在极端可行点上,因而我们可以从某个这种点上开始计算 H.接着我们按使 H 能取更小值的原则,来选取与这个极端可行点以直线联结的另一个端点上的极端可行点,进行交换.这种交换过程沿着边前进,一直持续到不再减少为止.这个交换算法称为单纯形方法,它的细节在下题中给出.

27.6 发展单纯形方法.

解 ☞ 令所讨论的问题是

$$0\leqslant x,\quad Ax\leqslant b,\quad H(x)=c^Tx=\text{极小值}.$$

我们首先介绍松弛变量(slack variable) $x_{n+1},\cdots,x_{n+m}$,使

$$\begin{aligned}
a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n+x_{n+1}&=b_1,\\
a_{21}x_1+a_{22}x_2+\cdots+a_{2n}x_n+x_{n+2}&=b_2,\\
\cdots\\
a_{m1}x_1+a_{m2}x_2+\cdots+a_{mn}x_n+x_{n+m}&=b_m.
\end{aligned}$$

注意这些松弛变量与其他 x_i 一样必须为非负的,松弛变量的使用允许我们以另外的方法对一个极端可行点加以确认.由于现在 $Ax\leqslant b$ 中等式的成立对应了一个松弛变量为零,一个极端可行点变成了在 $x_1,\cdots,x_{n+m}$ 中至少有 n 个变量为零的这样一个点.换言之,在一个极端可行点上这些变量中最多有 m 个非零,系数矩阵变成

$$\begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1n} & 1 & 0 & \cdots & 0\\
a_{21} & a_{22} & \cdots & a_{2n} & 0 & 1 & \cdots & 0\\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\
a_{m1} & a_{m2} & \cdots & a_{mn} & 0 & 0 & \cdots & 1
\end{pmatrix},$$

最后 m 列对应于松弛变量,令矩阵的列称作 $v_1,v_2,\cdots,v_{n+m}$.于是线性方程组可以写成

$$x_1v_1+x_2v_2+\cdots+x_{n+m}v_{n+m}=b.$$

现在假设我们知道一个极端可行点.为简单起见我们将取在这一点上的 $x_{m+1},\cdots,x_{m+n}$ 都为零所以 $x_1,\cdots,x_m$ 是(最多 m 个)非零变量.于是

$$x_1v_1+x_2v_2+\cdots+x_mv_m=b,\tag{1}$$

而相应的 H 值为

$$H_1=x_1c_1+x_2c_2+\cdots+x_mc_m.\tag{2}$$

假设向量 $v_1, \cdots, v_m$ 为线性独立的，所有 $n+m$ 个向量可以用这个基底来表示

$$v_j = v_{1j}v_1 + \cdots + v_{mj}v_m, \quad j = 1, \cdots, n+m. \tag{3}$$

同样定义

$$h_j = v_{1j}c_1 + \cdots + v_{mj}c_m - c_j, \quad j = 1, \cdots, n+m. \tag{4}$$

现在，假设我们尝试通过包含一部分 px_k 来还原 H_1，对 $k>m$ 及正的 p. 为保护约束当 $j=k$ 时我们以 p 乘(3)，p 还是待定的，并从(1)减去它，来得到

$$(x_1 - pv_{1k})v_1 + (x_2 - pv_{2k})v_2 + \cdots + (x_m - pv_{mk})v_m + pv_k = b.$$

类似地从(2)和(4)，新的 H 值将是

$$(x_1 - pv_{1k})c_1 + (x_2 - pv_{2k})c_2 + \cdots + (x_m - pv_{mk})c_m + pc_k = H_1 - ph_k.$$

这种变化仅当 $h_k>0$ 时将是有利的，在这种情况下最优的是使 p 尽可能地大而不让一个系数 $x_i - pv_{ik}$ 变成负的. 它提示选择

$$p = \min_i \frac{x_i}{v_{ik}} = \frac{x_l}{v_{lk}},$$

这最小值只是在具有正的 v_{ik} 项上取. 取这样选择的 p，c_l 的系数变成零，其他的为非负，而我们有一个新的极端可行点使 H 取值

$$H'_1 = H_1 - ph_k,$$

它肯定比 H_1 小. 我们也有一个新基底，改变基底向量 v_l 为新的 v_k. 该过程现在被重复直到所有 h_j 为负的，或是直到对某些正的 h_k 没有 v_{ik} 再是正的；在前面这种情况下眼前的极端可行点与任何邻近的极端点一样地好，这可以进一步地证明它与任何其他邻近的点一样地好或者不是. 在后一种情况下 p 可以任意地大因而没有关于 H 的极小值.

在做另一个交换之前，所有向量必须以新的基底来表示. 这类交换在我们矩阵求逆的那一节中已经做过，但是细节将被重复. 向量 v_l 要被向量 v_k 所替代. 从

$$v_k = v_{lk}v_1 + \cdots + v_{mk}v_m$$

我们解出 v_l 并代入(3)得到新的表达式

$$v_j = v'_{1j}v_1 + \cdots + v'_{l-1,j}v_{l-1} + v'_{kj}v_k + v'_{l+1,j}v_{l+1} + \cdots + v'_{mj}v_m,$$

其中

$$v'_{ij} = \begin{cases} v_{ij} - \dfrac{v_{lj}}{v_{lk}}v_{ik}, & \text{当 } i \neq l, \\ \dfrac{v_{lj}}{v_{lk}}, & \text{当 } i = l. \end{cases}$$

同时将 v_l 代入(1)得到

$$x'_1v_1 + \cdots + x'_{l-1}v_{l-1} + x'_kv_k + x'_{l+1}v_{l+1} + \cdots + x'_mv_m = b,$$

其中

$$x'_i = \begin{cases} x_i - \dfrac{x_l}{v_{lk}}v_{ik}, & \text{当 } i \neq l, \\ \dfrac{x_l}{v_{lk}}, & \text{当 } i = l. \end{cases}$$

此外，稍加计算证明

$$h'_j = v'_{1j}c_1 + \cdots + v'_{mj}c_m - c_j = h_j - \frac{v_{lj}}{v_{lk}}h_k,$$

而我们已有

$$H'_1 = H_1 - \frac{x_l}{v_{lk}}h_k.$$

方程组的这个完整集合可以通过展示不同的成份紧凑地概括如下：

$$\begin{bmatrix} x_1 & v_{11} & v_{12} & \cdots & v_{1,n+m} \\ x_2 & v_{21} & v_{22} & \cdots & v_{2,n+m} \\ & & \cdots\cdots & & \\ x_m & v_{m1} & v_{m2} & \cdots & v_{m,n+m} \\ H_1 & h_1 & h_2 & \cdots & h_{n+m} \end{bmatrix}$$

称 v_{ik} 为主元,在主元行上的各元素被主元来除,主元列的各元素除主元位置上为 1 其他均为零,而所有其他的元素服从于先前称之为矩形法则的,现在将以各种不同的例子来说明它.

27.7 用单纯形方法解题 27.1.

解 ☞ 在引进了松弛变量之后,约束方程变为

$$\begin{aligned} -x_1 + 2x_2 + x_3 &= 2, \\ x_1 + x_2 + x_4 &= 4, \\ x_1 + x_5 &= 3. \end{aligned}$$

所含的 5 个变量要求都是非负的,代替将 $x_2 - x_1$ 极大化我们将 $x_1 - x_2$ 极小化.这样的在极大和极小化问题之间的一个开关对我们来说是常常可用的.由于原点是一个极端可行点,我们可以选择 $x_1 = x_2 = 0, x_3 = 2, x_4 = 4, x_5 = 3$ 为出发点.这是非常方便的,由于它相当于选择 v_3, v_4 及 v_5 为我们的第一个基底,它使所有的 $v_{ij} = a_{ij}$.从此出发的阵式如下

基底	b	v_1	v_2	v_3	v_4	v_5
v_3	2	-1	②	1	0	0
v_4	4	1	1	0	1	0
v_5	3	1	0	0	0	1
	0	-1	1	0	0	0

与题 27.6 中的格式进行比较,人们发现这 6 个向量 b 及 $v_1, \cdots, v_5$ 形成顶部的三行而数 $H, h_1, \cdots, h_5$ 在底行中.只有 h_2 为正的.它确定了主元所在的列.在这一列有二个正的 v_{i2} 数,但是 2/2 比 4/1 小所以主元是 $v_{12} = 2$,这个数已被圈起.现在应用前题中的这些公式来产生一个新的阵式.顶行简单地被除以 2,而所有其他的元服从矩形法则.

基底	b	v_1	v_2	v_3	v_4	v_5
v_2	1	-1/2	1	1/2	0	0
v_4	3	3/2	0	-1/2	1	0
v_5	3	1	0	0	0	1
	-1	-1/2	0	-1/2	0	0

基底向量 v_3 为 v_2 所交换而所有向量现在均以这个新基底来表示。但是对这个例子更为重要的是,现在 h_j 无一为正的所以算法停止.$x_1 - x_2$ 的极小值是 -1(使 $x_2 - x_1$ 的极大值等于 1,如前).这个最小值在 $x_2 = 1, x_4 = 3, x_5 = 3$ 处实现,正如第一列所示.这些约束使 $x_1 = 0, x_3 = 0$,这是我们所预期的,由于 x_j 不对应于基底向量应恒为零.结果 $x_1 = 0, x_2 = 1$ 对应于我们前面的几何结论.注意单纯形算法曾将我们从可行点集合中的极端可行点(0, 0)带至极端点(0, 1),它证明为解点.(见图 27.1)

27.8 以单纯形方法解题 27.2.

解 ☞ 松弛变量及约束与前题中的相同.我们将 $H = -2x_1 - x_2$ 极小化.原点为一个极端点,我们可以从这个阵式开始:

基底	b	v_1	v_2	v_3	v_4	v_5
v_3	2	-1	2	1	0	0
v_4	4	1	1	0	1	0
v_5	3	①	0	0	0	1
	0	2	1	0	0	0

h_1 及 h_2 均为正的，所以我们有一种选择．选 $h_1=2$ 使 v_{13}为主元，由于3/1比4/1小．这个主元已被圈起，将 v_5 与 v_1 交换我们有一个新的基底，一个新的极端点，及一个新的阵式．

v_3	5	0	2	1	0	1
v_4	1	0	①	0	1	−1
v_1	3	1	0	0	0	1
	−6	0	1	0	0	−2

现在我们没有选择．新主元已被圈起并意味着，我们将 v_4 换成 v_2 具有下面的结果：

v_3	3	0	0	1	−2	3
v_2	1	0	1	0	1	−1
v_1	3	1	0	0	0	1
	−7	0	0	0	−1	−1

现在 h_j 无一为正，于是我们便停止．极小值为 −7．它与在题27.2中获得的 $2x_1+x_2$ 的极大值为7相一致，解点是在 $x_1=3,x_2=1$ 它也与题27.2所获得的结果相一致．单纯形方法曾将我们从(0,0)带至(3,0)又到(3,1)．其他对我们来说可用的选择在第一次交换时就会把我们带到以另外的方向围绕可行点集合．

27.9 以单纯形方法解题27.3．

解 ☞ 取松弛变量约束就变成

$$y_1 - y_2 - y_3 + y_4 = 1,$$
$$-2y_1 - y_2 + y_5 = -1.$$

所有的5个变量被要求为正的或为零．然而此时，原点($y_1=y_2=y_3=0$)不是一个可行点，如图27.2所示并且为执行过的负的 $y_5=-1$ 所证实的．因此我们不能跟随前两例中的出发过程，它建立在诸如

基底	b	v_1	v_2	v_3	v_4	v_5
v_4	1	1	−1	−1	1	0
v_5	−1	−2	−1	0	0	1

的一个阵式上．在 b 列中的负值 $y_5=-1$ 是不能被允许的，基本上我们的问题是我们没有一个极端可行点可以从它那儿开始出发的．寻找这样一个点的标准过程，甚至对比它大得多的问题，是介绍一个人工的基底．这里将第二个约束换成 $-2y_1-y_2+y_5-y_6=-1$ 就足够了，它包含了负的 b 分量，现在可以将一个新的列附加到我们前面的阵式中去．

基底	b	v_1	v_2	v_3	v_4	v_5	v_6
v_4	1	1	−1	−1	1	0	0
v_5	−1	−2	−1	0	0	1	−1

但是一个极端可行点现在对应于 $y_4=y_6=1$，所有其他 y_i 均为零，它使在这个基底中将 v_5 换成 v_6 变得自然，经过 v_6 行时只要求少数 n 个符号改变这个出发阵式的最后一行现在得到了解释．

基底	b	v_1	v_2	v_3	v_4	v_5	v_6
v_4	1	1	-1	1	1	0	0
v_6	1	②	1	0	0	-1	1
	w	$2w-2$	$w-4$	-3	0	$-w$	0

引进这个人工基底已经改变了我们的原始问题,除非我们可以肯定 y_6 最终证明是零,幸运地是它可以通过将要极小化的函数,$H=2y_1+4y_2+3y_3$,像它在题27.2中那样的,改变成 $H^*=2y_1+4y_2+3y_3+Wy_6$ 而得到安排.其中 W 是这样的一个大的正数,对一个极小化我们肯定必须使 y_6 等于零,以这些改变我们有 W 的一个出发 H 值.数 h_j 也可以得到计算而出发阵式的最后一行正如所示.

我们现在以正常的单纯形风格进行.由于 W 是大数且为正的,我们有两个正的 h_j 值的一个选择,选 h_1 导至加圈的主元.由于 v_6 没有进一步的意义,将 v_6 换成 v_1 带来一个新的阵式,由它最后一列就被摘除了.

v_4	$\frac{1}{2}$	0	$-\frac{3}{2}$	-1	1	$\frac{1}{2}$
v_1	$\frac{1}{2}$	1	$\frac{1}{2}$	0	0	$-\frac{1}{2}$
	1	0	-3	-3	0	-1

由于 h_j 无一为正,我们已经到达终点,极小值为 1,它与我们在题 27.3 中的几何的结论相符.此外,从第一列我们得到 $y_1=\frac{1}{2}$,$y_4=\frac{1}{2}$ 其他所有 y_j 等于零,它产生的极小点 $\left(\frac{1}{2},0,0\right)$ 在题 27.3 中也得到过.

27.10　极小化函数 $H=2y_1+4y_2+3y_3$ 受控于约束 $y_1-y_2-y_3\leqslant-2$,　$-2y_1-y_2\leqslant-1$,所有 y_j 为正或零.

解　松弛变量及人工基底将约束条件转化为

$$y_1-y_2-y_3+y_4-y_6=-2,$$
$$-2y_1-y_2+y_5-y_7=-1.$$

并且很像在前题中那样,我们很快便有这个出发阵式:

基底	b	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_6	2	-1	1	1	-1	0	1	0
v_7	1	2	①	0	0	-1	0	1
	$3W$	$W-2$	$2W-4$	$W-3$	$-W$	$-W$	0	0

要极小化的函数为

$$H^*=2y_1+4y_2+3y_3+Wy_6+Wy_7.$$

v_6	1	-3	0	①	-1	1	1
v_2	1	2	1	0	0	-1	0
	$W+4$	$-3W+6$	0	$W-3$	$-W$	$W-4$	0

而它决定这最后一行,对主元有不同的选择,我们选加圈的那个,它导致一个新的阵式通过将 v_7 换成 v_2 并将 v_7 这一列摘掉.一个新的主元也被加了圈于是跟之而来的最后的阵式为

v_3	1	-3	0	1	-1	1
v_2	1	2	1	0	0	-1
	7	-3	0	0	-3	-1

H^* 及 H 的最小值是 7，而它在(0,1,1)处出现.

对偶定理

27.11 什么是线性规划的对偶定理？

解 ☞ 考虑这两个线性规划问题：

问题 A	问题 B
$c^Tx=$极小，	$y^Tb=$极大，
$x\geqslant 0$，	$y\geqslant 0$，
$Ax\geqslant b$.	$y^TA\leqslant c^T$.

它们所以称为对偶问题，因为它们之间有许多关系，诸如下面的那些：

1. 若其中的一个问题有一个解，则另一个问题也如此，而且 c^Tx 的极小值与 y^Tb 的极大值相等.
2. 对其中的一个问题以通常的方法求得它的解向量，则对偶问题的解向量可以通过依次取松弛变量而得到，把零值分配给那些在最后基底中出现的，并给每一个其他的以相应的 $-h_j$ 值. 这里对这些结果不加证明而是使用我们前面的例子加以说明. 对偶性使得通过解 A 与 B 二个问题中的一个而得到两者的解成为可能.

27.12 证明题 27.1 及 27.3 为对偶题，并证明在题 27.11 中说明过的两个关系.

证 ☞ 包含了少许小小的改变. 将题 27.1 与 A 相配，我们以对 x_1-x_2 的极小化来代替对 x_2-x_1 的极大化，于是向量 c^T 为$(1,-1)$. 约束则改写为

$$x_1-2x_2\geqslant -2,\quad -x_1-x_2\geqslant -4,\quad -x_1\geqslant -3,$$

它使得

$$A=\begin{bmatrix}1 & -2\\ -1 & -1\\ -1 & 0\end{bmatrix},\quad b=\begin{bmatrix}-2\\ -4\\ -3\end{bmatrix}.$$

于是对问题 B 我们有

$$y^TA=\begin{bmatrix}y_1-y_2-y_3\\ -2y-y_2\end{bmatrix}\leqslant\begin{bmatrix}1\\ -1\end{bmatrix}.$$

它们是题 27.3 的约束. $y^Tb=$极大的条件也等价于

$$y^T(-b)=2y_1+4y_2+3y_3=\text{极小}.$$

所以题 27.3 与 B 也是匹配的，两个问题的极值证明为 1，它证明了题 27.11 的关系 1. 从题 27.7 中的最后的单纯形阵式我们得到 $x^T=(0,1)$与 $y^T=\left(\frac{1}{2},0,0\right)$，而从题 27.9 的计算中我们得到 $y^T=\left(\frac{1}{2},0,0\right)$ 与 $x^T=(0,1)$，证明了关系 2.

27.13 证明题 27.2 与 27.10 是对偶的.

证 ☞ 矩阵 A 及向量 b 与题 27.12 的相同. 然而，我们现在有 $c^T=(-2,-1)$. 它将题27.2与问题 A 相匹配，将题 27.10 与问题 B 相匹配，题 27.8 最后的阵式产生 $x^T=(3,1)$及 $y^T=(0,1,1)$，而同样的结果来自题 27.10. 这共同的 c^Tx 的极小与 y^Tb 的极大值是 -7.

27.14 证明如何使两人游戏等价于线性规划.

证 ☞ 令由正数 a_{ij}组成的付清矩阵(pay off 矩阵)是

$$A=\begin{bmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{bmatrix},$$

用它我们的意思是当游戏者 R 选了该矩阵的第 i 行而游戏者 C(独立地)选了 j 列,接下来 R 就要向 C 付清 a_{ij} 这个量.这就构成了游戏的一局.问题是要对每个游戏者在选择行或列时确定最好的策略.为更明确起见,令 C 选这三个列的概率分别为 p_1, p_2, p_3,则

$$p_1, p_2, p_3 \geqslant 0, 且 p_1 + p_2 + p_3 = 1$$

依赖于 R 对行的选择,C 现在有下面的三个量为他所期望赢的.

$$P_1 = a_{11}p_1 + a_{12}p_2 + a_{13}p_3,$$
$$P_2 = a_{21}p_1 + a_{22}p_2 + a_{23}p_3,$$
$$P_3 = a_{31}p_1 + a_{32}p_2 + a_{33}p_3.$$

令 P 为这三个数中最小的一个.于是,不论 R 怎么玩法,C 将有希望赢得的数至少在每一局上为 P,于是因此他问自己这个量 P 怎样才能极大化.由于全部所含的数都是正的,于是 P 也为正的,因而我们通过令

$$x_1 = \frac{p_1}{P}, \qquad x_2 = \frac{p_2}{P}, \qquad x_3 = \frac{p_3}{P},$$

得到一个等价的问题,并将 $F = x_1 + x_2 + x_3 = \frac{1}{P}$ 极小化.

不同的约束可以表示为 $x_1, x_2, x_3 \geqslant 0$ 且

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \geqslant 1,$$
$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \geqslant 1,$$
$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \geqslant 1.$$

这就是我们取 $c^T = b^T = (1,1,1)$ 对偶定理的类型 A 问题.

现在以 R 的观点来看事物.假设他选取 3 个行的概率分别为 q_1, q_2, q_3,依赖于 C 对列的选择,他有下面量中的一个作为它的期望损失:

$$q_1a_{11} + q_2a_{21} + q_3a_{31} \leqslant Q,$$
$$q_1a_{12} + q_2a_{22} + q_3a_{32} \leqslant Q,$$
$$q_1a_{13} + q_2a_{23} + q_3a_{33} \leqslant Q,$$

其中 Q 为三者中最大的一个.于是,不论 C 怎么玩法,R 将会有在每一局上将会有期望的损失不超过 Q.据此,他问怎样才能将这个量极小化.由于 $Q > 0$,我们令

$$y_1 = \frac{q_1}{Q}, \quad y_2 = \frac{q_2}{Q}, \quad y_3 = \frac{q_3}{Q},$$

并考虑等价的极大化问题

$$G = y_1 + y_2 + y_3 = \frac{1}{Q},$$

约束为 $y_1, y_2, y_3 \geqslant 0$ 且

$$y_1a_{11} + y_2a_{21} + y_3a_{31} \leqslant 1,$$
$$y_1a_{12} + y_2a_{22} + y_3a_{32} \leqslant 1,$$
$$y_1a_{13} + y_2a_{23} + y_3a_{33} \leqslant 1,$$

这就是我们取 $c^T = b^T = (1,1,1)$ 的对偶定理的类型 B 问题.我们已经发现了 R 问题与 C 问题是对偶的,这意味着极大值 P 与极小值 Q 将是相同的.故二个游戏者都在最优的平均付清上一致.它同时意味着对二个游戏者的最佳策略正好可以通过解同一个对偶问题找到,我们选择 R 的问题由于它避免了引进人工基底.

同样的论点应用于其他规模的付清矩阵.此外,所有 a_{ij} 是正的这个要求可以方便地去掉,由于若 a_{ij} 换成 $a_{ij} + a$,则 P 与 Q 便换成 $P + a$ 及 $Q + a$,因此改变的只是游戏的值而不是它的最佳策略.在下面将提供例题.

27.15 对具有矩阵

$$A = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 1 & 2 & 0 \end{bmatrix}$$

的游戏,寻找对二个游戏者的最佳策略以及最佳付清.

解 ☞ 我们代之以在约束条件

$$\begin{aligned} y_2+y_3+y_4 \quad &= 1, \\ y_1 \quad +2y_3 \quad +y_5 &= 1, \\ 2y_1+y_2 \quad +y_6 &= 1. \end{aligned}$$

下函数 $-G=-y_1-y_2-y_3$ 的极小化,所有 y_j 包括松弛变量 y_4, y_5, y_6 都是非负的.由于原点是一个极端可行点,我们的出发阵式为

基底	b	v_1	v_2	v_3	v_4	v_5	v_6
v_4	1	0	1	1	1	0	0
v_5	1	1	0	2	0	1	0
v_6	1	②	1	0	0	0	1
	0	1	1	1	0	0	0

使用标明的主元我们作三次交换如下:

v_4	1	0	1	1	1	0	0
v_5	1/2	0	-1/2	②	0	1	-1/2
v_1	1/2	1	1/2	0	0	0	1/2
	-1/2	0	1/2	1	0	0	-1/2

v_4	3/4	0	$\frac{5}{4}$	0	1	-1/2	$\frac{1}{4}$
v_3	1/4	0	$-\frac{1}{4}$	1	0	1/2	-1/4
v_1	1/2	1	1/2	0	0	0	1/2
	$-\frac{3}{4}$	0	3/4	0	0	-1/2	-1/4

v_2	3/5	—	—	—	—	—	—
v_3	2/5	—	—	—	—	—	—
v_1	1/5	—	—	—	—	—	—
	-6/5	0	0	0	-3/5	-1/5	-2/5

从最后的阵式我们推导最佳付清,或游戏的值,是$\frac{5}{6}$. R 的最优策略可以直接地通过规范解 $y_1=\frac{1}{5}, y_2=\frac{3}{5}, y_3=\frac{2}{5}$ 得到,概率 q_1, q_2, q_3 必须与这些 y_j 成比例,但是其和必须为 1,据此

$$q_1=\frac{1}{6}, \quad q_2=\frac{3}{6}, \quad q_3=\frac{2}{6}.$$

为了得到关于 C 的最优策略我们注意到在最后的基底中没有松弛变量,于是将 $-h_j$ 放到(非基底)松弛变量的位置上

$$x_1=\frac{3}{5}, \qquad x_2=\frac{1}{5}, \qquad x_3=\frac{2}{5}.$$

规范化带来

$$p_1=\frac{3}{6}, \quad p_2=\frac{1}{6}, \quad p_3=\frac{2}{6}.$$

假如游戏者之一用这个最优策略来混合他的选择,平均付清将是$\frac{5}{6}$,为使游戏公平起见,所有付清

可以按这个量来减少,或者 C 可以在每局结束以前要求付这个量.

27.16 对具有矩阵

$$A = \begin{bmatrix} 0 & 3 & 4 \\ 1 & 2 & 1 \\ 4 & 3 & 0 \end{bmatrix}$$

的游戏,对每个游戏者找出最优策略以及最优付清.

解 ☞ 注意中心的元素既是行的极大也是列的极小,它也是行的极大中最小的,也是列的极小中最大的.这样的一个鞍点恒等于一个具有纯策略(pure strategies)的游戏.单纯形直接引出这个使用鞍点为主元的结果出发的阵式如下:

基底	b	v_1	v_2	v_3	v_4	v_5	v_6
v_4	1	0	1	4	1	0	0
v_5	1	3	②	3	0	1	0
v_6	1	4	1	0	0	0	1
	0	1	1	1	0	0	0

一次交换就足够了:

v_4	$\frac{1}{2}$	—	—	—	—	—	—
v_2	$\frac{1}{2}$	—	—	—	—	—	—
v_6	$\frac{1}{2}$	—	—	—	—	—	—
	$-\frac{1}{2}$	$-\frac{1}{2}$	0	$-\frac{1}{2}$	0	$-\frac{1}{2}$	0

最优的付清是 $-\frac{1}{2}$ 的负倒数,即是主元元素 2. 直接获得对 R 的最优策略,由于 $y_1 = 0, y_2 = \frac{1}{2}, y_3 = 0$,我们规范化去获得纯策略

$$q_1 = 0, \qquad q_2 = 1, \qquad q_3 = 0.$$

只有第二行应一直被使用.关于 C 的策略通过松弛变量得到,由于 v_4 及 v_6 在最后的基底中我们有 $x_1 = x_3 = 0$,以及最后 $x_2 = -h_5 = 1/2$.规范化,我们获另外的纯策略

$$p_1 = 0, \qquad p_2 = 1, \qquad p_3 = 0.$$

补 充 题

27.17 作一个图示说明同时满足下面约束的所有的点

$$0 \leqslant x_1, \quad 0 \leqslant x_2, \quad x_1 + 2x_2 \leqslant 4, \quad -x_1 + x_2 \leqslant 1,$$
$$x_1 + x_2 \leqslant 3.$$

27.18 前题中的 5 个极端可行点是什么? 在哪个极端可行点上 $F = x_1 - 2x_2$ 取它的极小值又极小值为何? 在哪个极端可行点上这个函数取它的极大值?

27.19 通过应用单纯形方法在题 27.17 约束的限制下找出 $F = x_1 - 2x_2$ 的极小值,当你用几何方法时得到的值是否相同? 极端可行点是否相同?

27.20 什么是题 27.19 的对偶问题? 证明,通过使用该题中获得的最后单纯形阵式,得到对偶问题的解向量 $y_1 = \frac{1}{3}, y_2 = \frac{4}{3}, y_3 = 0$.

27.21 通过应用单纯形方法在题 27.17 约束的限制下求 $F = x_1 - 2x_2$ 的最大值.(极小化 $-F$)用几何方法

你得到的结果是否相同?

27.22 什么是题 27.21 的对偶问题? 从该题的最后单纯形阵式中得到它的解.

27.23 直接由单纯形方法解题 27.19 的对偶问题,使用一个外加的变量于一个人工基底,约束该读作

$$-y_1+y_2-y_3+y_4=1,$$
$$-2y_1-y_2-y_3+y_5-y_6=-2,$$

以 y_4 与 y_5 为松弛变量. $H=4y_1+y_2+3y_3$ 为要极小化的函数,由最后的阵式重新找到对偶问题的解及题 27.19 它本身的解.

27.24 极小化 $F=2x_1+x_2$,在约束

$$3x_1+x_2\geqslant 3,\quad 4x_1+3x_3\geqslant 6,\quad x_1+2x_2\geqslant 2$$

的限制下所有 x_j 是非负的(得到的解为 $x_1=3/5,x_2=6/5$).

27.25 几何地说明关于一个在题 27.17 约束限制下 $F=x_1-x_2$ 的极小有无穷多个解点.它们在何处? 证明单纯形方法直接地产生一个极端解点,而它还产生另外的,假如作了最后的一个 v_3 与 v_1 的交换,即使相应的 h_j 值为零.解点的集合是连结这些极端点的线段.

27.26 在约束

$$2x_1+2x_2+x_3\leqslant 7,\quad x_2+x_4\geqslant 1,$$
$$2x_1+x_2+2x_3\leqslant 4,\quad x_2+x_3+x_4=3.$$

的限制下极小化 $F=x_1+x_4$,所有 x_j 为非负的.(极小值为零,在多于一个的可行点上出现.)

27.27 对游戏

$$A=\begin{bmatrix}1&3\\4&2\end{bmatrix}$$

使用单纯形方法寻找最优策略及付清. $\left[\text{付清是 } 2.5, R \text{ 的策略是}(1/2,1/2)\text{而对 } C \text{ 来说是}\left(\frac{1}{4},\frac{3}{4}\right).\right]$

27.28 解具有矩阵

$$A=\begin{bmatrix}0&3&-4\\3&0&5\\-4&5&0\end{bmatrix}$$

的游戏,证明最优付清为 $\frac{10}{7}$,对 R 的最佳策略为 $\left(\frac{5}{14},\frac{4}{7},\frac{1}{14}\right)$.而对 C 的是相同的.

27.29 由单纯形方法解下面的游戏

$$A=\begin{bmatrix}0&0&1&1\\1&1&-2&-2\\1&-2&1&-2\\-2&3&-2&3\end{bmatrix}.$$

27.30 对下面的函数求极小-极大二次多项式.什么是极小-极大误差以及它在何处达到?

x	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
$y(x)$	5	5	4	2	1	3	7	10	12

27.31 求关于

$$y(x)=\frac{1}{1+(4.1163x)2}\qquad x=0(0.01)1$$

的极小-极大二次多项式以及极小-极大误差以及达到它的自变量.

27.32 对前题的函数寻找一个三次逼近,它的结果是什么? 又怎样从前题的结果来预报它?

27.33 在约束

$$x_1+x_2+3x_3+x_4\leqslant 5$$
$$x_1+x_3-4x_4\leqslant 2$$

以及所有 $x_k\geqslant 0$ 的限制下极小化 $x_1-x_2+2x_3$.

27.34　解前题的对偶问题.

27.35　在约束

$$x_1 - x_2 \leqslant 2, \quad x_1 + x_2 \leqslant 6, \quad x_1 + 2x_2 \leqslant A$$

的限制下及所有 $x_k \geqslant 0$ 极小化 $2x_1 + x_2$，处理 $A = 0, 3, 6, 9, 12$ 的情况.

27.36　利用线性规划在下面的游戏中寻找对两个游戏者都是最优的策略.

$$\begin{bmatrix} -6 & 4 \\ 4 & -2 \end{bmatrix}$$

27.37　作为线性规划来解具有付清矩阵为

$$\begin{bmatrix} 3 & 1 \\ 2 & 3 \end{bmatrix}$$

的游戏.

第二十八章　超定方程组

问题的性质

一个超定线性方程组其形式为

$$Ax = b,$$

矩阵 A 的行比列多.正常情况不存在解向量,因而像这样写出的方程是没有意义的.这个方程组也称为不相容的.在实验或计算工作中,当在精度可达到的情况下产生的结果比所需要的更多时超定方程组会出现.在某种意义上说,一堆不精确的、冲突的信息变成欠完美的结果的替代物,而人们希望以某种方式从矛盾中尽力获取对精确结果的一个好的逼近.

两种处理方法

这两种主要方法包含残余向量

$$R = Ax - b.$$

由于 R 通常不会化为零向量,努力以这样一种方式选取 x 使得 R[*] 在某种意义下被极小化.

1. 一个超定方程组的**最小二乘解**是指向量 x,它使残余向量的各分量的平方和为极小.以向量的语言我们要

   $$R^T R = \text{极小}.$$

   对于 m 个方程 n 个未知量,具有 $m > n$,在第 21 章中用过的那一类的讨论可引出法方程

   $$(a_1, a_1)x_1 + \cdots + (a_1, a_n)x_n = (a_1, b),$$
   $$\cdots$$
   $$(a_n, a_1)x_1 + \cdots + (a_n, a_n)x_n = (a_n, b).$$

   它决定 x 的分量.这里

   $$(a_i, a_j) = a_{1i}a_{1j} + \cdots + a_{mi}a_{mj}$$

   是 A 的二个列向量的标量积.

2. **Chebyshev 或极小化极大解**是指向量 x,它使残余向量绝对值最大的分量为极小.也就是说,我们尝试极小化

   $$r = \max(|r_1|, \cdots, |r_m|)$$

   其中 r_i 为 R 的分量.当 $m = 3$, $n = 2$ 时它转变成一组约束含有要加以极小化的 r:

   $$a_{11}x_1 + a_{12}x_2 - b_1 \leqslant r, \quad -a_{11}x_1 - a_{12}x_2 + b_1 \leqslant r,$$
   $$a_{21}x_1 + a_{22}x_2 - b_2 \leqslant r, \quad -a_{21}x_1 \quad a_{22}x_2 + b_2 \leqslant r,$$
   $$a_{31}x_1 + a_{32}x_2 - b_3 \leqslant r, \quad -a_{31}x_1 - a_{32}x_2 + b_3 \leqslant r.$$

   这容易转化成一个线性规划问题.类似的线性规划解 m 和 n 为任意的情况.

题　　解

最小二乘解

28.1　导出法方程用于寻求超定线性方程组的最小二乘解.

---

* 译注:原文为 r.

解 ☞ 令已知方程组为

$$a_{11}x_1 + a_{12}x_2 = b_1,$$
$$a_{21}x_1 + a_{22}x_2 = b_2,$$
$$a_{31}x_1 + a_{32}x_2 = b_3,$$

它只包含了二个未知量而且只是轻微超定，但是其细节对更大的方程组是几乎一样的.正常情况我们不可能满足我们方程组中所有的三个方程.这个问题就像它现在的这个样子很可能没有解的.据此，我们将它改写为

$$a_{11}x_1 + a_{12}x_2 - b_1 = r_1,$$
$$a_{21}x_1 + a_{22}x_2 - b_2 = r_2,$$
$$a_{31}x_1 + a_{32}x_2 - b_3 = r_3.$$

其中数 r_1, r_2, r_3 称为残量，并且寻找数 x_1, x_2 使 $r_1^2 + r_2^2 + r_3^2$ 为极小.因为

$$\begin{aligned} r_1^2 + r_2^2 + r_3^2 = &(a_{11}^2 + a_{21}^2 + a_{31}^2)x_1^2 + (a_{12}^2 + a_{22}^2 + a_{32}^2)x_2^2 \\ &+ 2(a_{11}a_{12} + a_{21}a_{22} - a_{31}a_{32})x_1x_2 \\ &- (a_{11}b_1 + a_{21}b_2 + a_{31}b_3)x_1 \\ &- 2(a_{12}b_1 + a_{22}b_2 + a_{32}b_3)x_2 + (b_1^2 + b_2^2 + b_3^2), \end{aligned}$$

令它对 x_1 及 x_2 的导数为零得出的结果为一对法方程

$$(a_1, a_1)x_1 + (a_1, a_2)x_2 = (a_1, b),$$
$$(a_2, a_1)x_1 + (a_2, a_2)x_2 = (a_2, b).$$

该式中的圆括弧表示

$$(a_1, a_1) = a_{11}^2 + a_{21}^2 + a_{31}^2,$$
$$(a_1, a_2) = a_{11}a_{12} + a_{21}a_{22} + a_{31}a_{32},$$

等等，这些是原方程中各列系数之间的标量积.对 m 个方程 n 个未知量($m > n$)的一般问题

$$a_{11}x_1 + \cdots + a_{1n}x_n = b_1,$$
$$a_{21}x_1 + \cdots + a_{2n}x_n = b_2,$$
$$\cdots\cdots\cdots\cdots$$
$$a_{m1}x_1 + \cdots + a_{mn}x_n = b_m.$$

一个几乎完全一样的讨论引出法方程

$$(a_1, a_1)x_1 + (a_1, a_2)x_2 + \cdots + (a_1, a_n)x_n = (a_1, b),$$
$$(a_2, a_1)x_1 + (a_2, a_2)x_2 + \cdots + (a_2, a_n)x_n = (a_2, b),$$
$$\cdots\cdots\cdots\cdots$$
$$(a_n, a_1)x_1 + (a_n, a_2)x_2 + \cdots + (a_n, a_n)x_n = (a_n, b).$$

这是一个对称、正定方程组.

同时值得注意的是眼前的这个问题又一次地与题 21.7 及 21.8 中的一般最小二乘近似模型相符合.刚才所得到的结果作为一个特例马上能得到，其中向量空间 E 由 m 维向量所组成，例如，矩阵 A 的列向量(我们用 $a_1, a_2, \cdots, a_n$ 来表示)以及数 b_i 的列(我们用 b 来表示).子空间 S 就是矩阵 A 的域(range)，也就是指向量 Ax 的集合，我们在 S 中要找一个向量 p，它极小化

$$\| p - b \|^2 = \| Ax - b \|^2 = \sum r_i^2,$$

而这个向量是 b 在 S 上的正交投影，由 $(p - b, u_k) = 0$ 所决定，其中 u_k 是关于 S 的某个基底.将这个基底选成 $u_k = a_k, k = 1, \cdots, n$，我们有通常的表达式 $p = x_1a_1 + \cdots + x_na_n$(记号与我们一般模型的稍有不同)，代入后便引出法方程.

28.2 找出方程组

$$x_1 - x_2 = 2,$$
$$x_1 + x_2 = 4,$$
$$2x_1 + x_2 = 8.$$

的最小二乘解.

解 ☞ 形成所需要的标量积，我们得到法方程

$$6x_1+2x_2=22,\quad 2x_1+3x_2=10.$$

由此得 $x_1=\frac{23}{7}$ 及 $x_2=\frac{8}{7}$.对应于这个 x_1 及 x_2 的残量为 $r_1=\frac{1}{7}$, $r_2-\frac{3}{7}$, 及 $r_3=-\frac{2}{7}$,而它们的平方和为$\frac{2}{7}$.因此标准差为 $\rho=\sqrt{\frac{2}{21}}$.这比作其他选择的 x_1 及 x_2 都要小.

28.3　假设在题 28.2 已经是超定的方程组上再加上三个方程

$$x_1+2x_2=4,$$
$$2x_1-x_2=5,$$
$$x_1-2x_2=2$$

找出这个 6 个方程的方程组的最小二乘解.

解 ☞ 仍然形成标量积我们得到法方程 $12x_1=38$, $12x_2=9$,解之得 $x_1-\frac{19}{6}$, $x_2=\frac{3}{4}$. 6 个残量为 5, −1, −11, 8, 7 及 −4,所有的都要除以 12.标准差为 $\rho=\sqrt{\frac{23}{72}}$.

28.4　在大方程组的情况下,法方程组可以如何去解?

解 ☞ 由于法方程组为对称正定的,有几种方法完成得非常好.可以应用 Gauss 消去法.如果它的主元是通过递减主对角元来选取,则它能保持对称性直到最后.因而可省去几乎一半的计算量.

Chebyshev 解

28.5　阐明超定线性方程组的 Chebyshev 解如何能通过线性规划方法求得.

解 ☞ 我们再次处理题 28.1 的小方程组,对于大的方程组其细节几乎完全一样.令 r 为诸残量之绝对值中最大者,故有 $|r_1|\leqslant r$, $|r_2|\leqslant r$, $|r_3|\leqslant r$.这意味着 $r_1\leqslant r$, $-r_1\leqslant r$,对 r_2 及 r_3 的要求相同.回顾残量的定义我们现在有 6 个不等式:

$$a_{11}x_1+a_{12}x_2-b_1\leqslant r,\quad -a_{11}x_1-a_{12}x_2+b_1\leqslant r,$$
$$a_{21}x_1+a_{22}x_2-b_2\leqslant r,\quad -a_{21}x_1-a_{22}x_2+b_2\leqslant r,$$
$$a_{31}x_1+a_{32}x_2-b_3\leqslant r,\quad -a_{31}x_1-a_{32}x_2+b_3\leqslant r.$$

假如我们还假设 x_1 及 x_2 必须为非负的,并记住 Chebyshev 解是定义为选择 x_1, x_2 使 r 极小,于是我们显然有一个线性规划问题.为方便起见将它稍加改变.以 r 通除并令 $x_1/r=y_1$, $x_2/r=y_2$, $1/r=y_3$.约束就变成了

$$a_{11}y_1+a_{12}y_2-b_1y_3\leqslant 1,\qquad -a_{11}y_1-a_{12}y_2+b_1y_3\leqslant 1,$$
$$a_{21}y_1+a_{22}y_2-b_2y_3\leqslant 1,\qquad -a_{21}y_1-a_{22}y_2+b_2y_3\leqslant 1,$$
$$a_{31}y_1+a_{32}y_2-b_3y_3\leqslant 1,\qquad -a_{31}y_1-a_{32}y_2+b_3y_3\leqslant 1.$$

因而我们必须极大化 y_3.或者,同一件事,使 $F=-y_3$ 极小.线性规划问题可以直接地由原来的超定方程组形成.推广到大的方程组几乎是显然的事.x_j 为正的这一条件在实践中是常常满足的,这些数表示长度或其他的物理度量.如果不能被满足,那么可以做一个平移 $z_j=x_j+c$,或是可以将线性规划算法加以修改.

28.6　应用线性规划方法求题 28.2 中方程组的 Chebyshev 解.

解 ☞ 在每个约束上加一个松弛变量,我们有

$$\begin{aligned}
y_1-y_2-2y_3+y_4 &= 1,\\
y_1+y_2+4y_3+y_5 &= 1,\\
2y_1+y_2-8y_3+y_6 &= 1,\\
-y_1+y_2+2y_3+y_7 &= 1,\\
-y_1-y_2+4y_3+y_8 &= 1,\\
-2y_1-y_2+8y_3+y_9 &= 1.
\end{aligned}$$

加上 $F=-y_3$ 要被极小化以及所有 y_i 为非负的.出发阵式以及按单纯形算法所作的二次交换如图 28.1 所示.对应于松弛变量的 6 列被省略,因为它们事实上并不包含关键性信息.从最后的阵式我们得到 y_1-10 及 $y_2=y_3=3$.这使得 $r=1/y_3=\frac{1}{3}$ 以及 $x_1=\frac{10}{3}$, $x_2=1$.三个残量为$\frac{1}{3}$, $\frac{1}{3}$, $-\frac{1}{3}$,所以

熟悉的 Chebyshev 的等误差大小性质又一次出现.

基底	b	v_1	v_2	v_3
v_4	1	1	−1	−2
v_5	1	1	1	−4
v_6	1	2	1	−8
v_7	1	−1	1	2
v_8	1	−1	−1	4
v_9	1	−2	−1	⑧
	0	0	0	1

基底	b	v_1	v_2	v_3
v_4	$\frac{5}{4}$	$(\frac{1}{2})$	$-\frac{5}{4}$	0
v_5	$\frac{3}{2}$	0	$\frac{1}{2}$	0
v_6	2	0	0	0
v_7	$\frac{3}{4}$	$-\frac{1}{2}$	$\frac{5}{4}$	0
v_8	$\frac{1}{2}$	0	$-\frac{1}{2}$	0
v_3	$\frac{1}{8}$	$-\frac{1}{4}$	$-\frac{1}{8}$	1
	$-\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	0

基底	b	v_1	v_2	v_3
v_1	$\frac{5}{2}$	1	$-\frac{5}{2}$	0
v_5	$\frac{3}{2}$	0	$(\frac{1}{2})$	0
v_6	2	0	0	0
v_7	2	0	0	0
v_8	$\frac{1}{2}$	0	$-\frac{1}{2}$	0
v_3	$\frac{3}{4}$	0	$-\frac{3}{4}$	1
	$-\frac{3}{4}$	0	$\frac{3}{4}$	0

基底	b	v_1	v_2	v_3
v_1	10	1	0	0
v_2	3	0	1	0
v_6	2	0	0	0
v_7	2	0	0	0
v_8	2	0	0	0
v_3	3	0	0	1
	−3	0	0	0

图 28.1

28.7 应用线性规划方法求题 28.3 中超定方程组的 Chebyshev 解.

解 ☞ 6 个附加的约束带来外加的 6 个松弛变量 $y_{10},\cdots,y_{15}$. 细节与题 28.6 的十分类似,在图 28.2 中再次略去松弛变量列,它概括了单纯形算法的 3 次交换. 在最后一次交换后我们得到 $y_1=\frac{13}{3}$, $y_2=1$, $y_3=\frac{4}{3}$. 于是 $r=\frac{3}{4}$ 与 $x_1=\frac{13}{4}$, $x_2=\frac{3}{4}$. 6 个残量为 2, 0, −3, 3, 3 及 −1, 所有的都要除以 4. 再次有 3 个残量等于极小化极大残量 r,现在其余的是更小的. 在一般问题中 $n+1$ 个等残量,其余的是更小的,与 Chebyshev 解一致,n 是未知量的个数.

28.8 将最小二乘解与 Chebyshev 解的残量进行比较.

解 ☞ 对一个数 $x_1,\cdots,x_n$ 的任意集合,令 $|r|_{\max}$ 表绝对值最大的残量. 则 $r_1^2+\cdots+r_m^2\leqslant m|r|_{\max}^2$ 因而标准差肯定不超过 $|r|_{\max}$. 但是最小二乘解是所有中最小的标准差,因而,以 ρ 表这个误差,$\rho\leqslant|r|_{\max}$. 特别当 x_j 为 Chebyshev 解时这不等式成立,在这种情况下 $|r|_{\max}$ 就是我们曾经称之为 r 的那个. 但是 Chebyshev 解也有它的最大误差为最小的性质,所以假如 $|\rho|_{\max}$ 表示最小二乘解的绝对值最大的残量,则 $r|_{\max}\leqslant|\rho|_{\max}$. 将这二个不等式合在一起,$\rho\leqslant r\leqslant|\rho|_{\max}$,于是我们有 Chebyshev 误差两侧都有界. 由于最小二乘解常常容易被找到,这最后的结果可以用来决定是否值得继续往下进行,来得到,由 Chebyshev 解所带来的最大残量的进一步减少.

28.9 应用上题于 28.2 的方程组.

解 ☞ 我们已经得到 $\rho=\sqrt{\frac{2}{21}}$, $r=\frac{1}{3}$ 及 $|\rho|_{\max}=3/7$ 这确如题 28.8 所提示的,它们正是稳定地增加的. 最小二乘残量之一为其他的 3 倍大,这事实已经推荐了对 Chebyshev 解的寻查.

28.10 应用题 28.8 于题 28.3 的方程组.

解 ☞ 我们已经得到 $\rho=\sqrt{\frac{23}{72}}$, $r=\frac{3}{4}$ 及 $|\rho|_{\max}=\frac{11}{12}$,这种布局确实支持了对 Chebyshev 解的寻查.

基底	b	v_1	v_2	v_3
v_4	1	1	−1	−2
v_5	1	1	1	−4
v_6	1	2	1	−8
v_7	1	−1	1	2
v_8	1	−1	−1	4
v_9	1	−2	−1	$\circled{8}$
v_{10}	1	1	2	−4
v_{11}	1	2	−1	−5
v_{12}	1	1	−2	−2
v_{13}	1	−1	−2	4
v_{14}	1	−2	1	5
v_{15}	1	−1	2	2
	0	0	0	1

基底	b	v_1	v_2	v_3
v_4	$\frac{5}{4}$	$\frac{1}{2}$	$-\frac{5}{4}$	0
v_5	$\frac{3}{2}$	0	$\frac{1}{2}$	0
v_6	2	0	0	0
v_7	$\frac{3}{4}$	$-\frac{1}{2}$	$\frac{5}{4}$	0
v_8	$\frac{1}{2}$	0	$-\frac{1}{2}$	0
v_3	$\frac{1}{8}$	$-\frac{1}{4}$	$-\frac{1}{8}$	1
v_{10}	$\frac{3}{2}$	0	$\frac{3}{2}$	0
v_{11}	$\frac{13}{8}$	$\circled{\frac{3}{4}}$	$-\frac{13}{8}$	0
v_{12}	$\frac{5}{4}$	$\frac{1}{2}$	$-\frac{9}{4}$	0
v_{13}	$\frac{1}{2}$	0	$-\frac{3}{2}$	0
v_{14}	$\frac{3}{8}$	$-\frac{3}{4}$	$\frac{13}{8}$	0
v_{15}	$\frac{3}{4}$	$-\frac{1}{2}$	$\frac{9}{4}$	0
	$-\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	0

基底	b	v_1	v_2	v_3
v_4	$\frac{1}{6}$	0	$-\frac{1}{6}$	0
v_5	$\frac{3}{2}$	0	$\frac{1}{2}$	0
v_6	2	0	0	0
v_7	$\frac{11}{6}$	0	$\frac{1}{6}$	0
v_8	$\frac{1}{2}$	0	$-\frac{1}{2}$	0
v_3	$\frac{2}{3}$	0	$-\frac{2}{3}$	1
v_{10}	$\frac{3}{2}$	0	$\circled{\frac{3}{2}}$	0
v_1	$\frac{13}{6}$	1	$-\frac{13}{6}$	0
v_{12}	$\frac{1}{6}$	0	$-\frac{7}{6}$	0
v_{13}	$\frac{1}{2}$	0	$-\frac{3}{2}$	0
v_{14}	2	0	0	0
v_{15}	$\frac{11}{6}$	0	$\frac{7}{6}$	0
	$-\frac{2}{3}$	0	$\frac{2}{3}$	0

基底	b	v_1	v_2	v_3
v_4	$\frac{1}{3}$	0	0	0
v_5	1	0	0	0
v_6	2	0	0	0
v_7	$\frac{5}{3}$	0	0	0
v_8	1	0	0	0
v_3	$\frac{4}{3}$	0	0	1
v_2	1	0	1	0
v_1	$\frac{13}{3}$	1	0	0
v_{12}	$\frac{4}{3}$	0	0	0
v_{13}	2	0	0	0
v_{14}	2	0	0	0
v_{15}	$\frac{2}{3}$	0	0	0
	$-\frac{4}{3}$	0	0	0

图 28.2

补 充 题

28.11 对方程组

$$x_1 - x_2 = -1, 2x_1 - x_2 = 2,$$
$$x_1 + x_2 = 8, \qquad 2x_1 + x_2 = 14.$$

求最小二乘解.计算这解的标准差.

28.12 对题 28.11 中求得的解比较 $|\rho|_{\max}$ 与 ρ.

28.13 寻找题 28.11 中方程组的 Chebyshev 解并比较它的 r 值与 ρ 及 $|\rho|_{\max}$.

28.14 对方程组

$$x_1 + x_2 - x_3 = 5, \qquad x_1 + 2x_2 - 2x_3 = 1,$$
$$2x_1 - 3x_2 + x_3 = -4, \qquad 4x_1 - x_2 - x_3 = 6.$$

寻找最小二乘解及 Chebyshev 解.

28.15 假设,已知 $-1 \leqslant x_j$.寻找下面方程组的 Chebyshev 解.通过首先令 $z_j = x_j + 1$,它保证了 $0 \leqslant z_j$.同时求最小二乘解.

$$2x_1 - 2x_2 + x_3 + 2x_4 = 1 \qquad -2x_1 - 2x_2 + 3x_3 + 3x_4 = 4$$
$$x_1 + x_2 + 2x_3 + 4x_4 = 1 \qquad -x_1 - 3x_2 - 3x_3 + x_4 = 3$$
$$x_1 - 3x_2 + x_3 + 2x_4 = 2 \qquad 2x_1 + 4x_2 + x_3 + 5x_4 = 0$$

28.16 寻找方程组

$$x_1 = 0, \qquad x_1 + x_2 = -1,$$
$$x_2 = 0, \qquad 0.1x_1 + 0.1x_2 = 0.1.$$

的最小二乘解.标准差是什么?

28.17 寻找题 28.16 中方程组的 Chebyshev 解.

28.18 测量所得的 4 个高度 x_1, x_2, x_3, x_4,连同高度的 6 个差一起,如下所示.寻找最小二乘解.

$$x_1 = 3.47, \quad x_2 = 2.01, \quad x_3 = 1.58, \quad x_4 = 0.43,$$
$$x_1 - x_2 = 1.42, \quad x_1 - x_3 = 1.92, \quad x_1 - x_4 = 3.06,$$
$$x_2 - x_3 = 0.44, \quad x_2 - x_4 = 1.53, \quad x_3 - x_4 = 1.20.$$

28.19 一个量 x 被测量了 N 次,其结果是 $a_1, a_2, \cdots, a_N$.以最小二乘法解超定方程组

$$x = a_i, \qquad i = 1,2,\cdots,N,$$

什么 x 值会出现?

28.20 测量二个量 x 及 y,连同它们的差 $x-y$ 以及和 $x+y$.

$$x = A, \quad y = B, \quad x - y = C, \quad x + y = D.$$

以最小二乘法解超定方程组.

28.21 三角形的三个角测得为 A_1, A_2, A_3.若以 x_1, x_2, x_3 表示准确值,我们被引至超定方程组

$$x_1 = A_1, \quad x_2 = A_2, \quad \pi - x_1 - x_2 = A_3.$$

以最小二乘法解它.

28.22 直角三角形的两条边测得为 A 与 B 而斜边为 C.令 L_1, L_2 及 H 表示精确值,并令 $x_1 = L_1^2, x_2 = L_2^2$.考虑超定方程组

$$x_1 = A^2, \quad x_2 = B^2, \quad x_1 + x_2 = C^2,$$

并得到 x_1 与 x_2 的最小二乘估计值,从它们来估计 L_1, L_2,及 H.

28.23 证明 $Ax = b$ 的最小二乘解的法方程等价于 $A^TAx = A^Tb$ *.

---

* 译注:原文为 $A^TA = A^Tb$.

第二十九章 边值问题

问题的性质

这是一个流行很广很深的课题,它的变化和算法可以写几本书.本章仅提供许多已实现的想法的实例.这就意味着难免涉及的范围较小,但忽略它却是完全不可取的.

边值问题是对区域 R 内的微分方程或方程组求解,这些解在 R 的边界上满足附加条件.实际应用产生了大量这类问题.经典的常微分方程两点边值问题包括一个二阶方程,一个初始条件和一个终点条件

$$y'' = f(x, y, y'), \quad y(a) = A, \quad y(b) = B.$$

这里的区域 R 是区间(a, b),边界有两点组成.典型的偏微分方程问题是 Dirichlet 问题.要求 Laplace 方程

$$U_{xx} + U_{yy} = 0$$

在 xy 平面的区域R 内部成立.在 R 的边界$U(x, y)$取指定值.这两个例子给出了两类重要的边值问题.

题解

1. **叠加原理**对线性问题很有用.例如求解

$$y'' = q(x)y, \qquad y(a) = A, \qquad y(b) = B,$$

我们可以用第十九章的办法.解两个初值问题

$$y_1'' = q(x)y_1, \qquad y_1(a) = 1, \qquad y_1(b) = 0,$$
$$y_2'' = q(x)y_2, \qquad y_2(a) = 0, \qquad y_2(b) = 1.$$

然后就有

$$y(x) = Ay_1(x) + By_2(x).$$

2. 当问题是线性时,**用矩阵问题来代替**也是一种选择.例如用二阶差商代替 $y''(x_k)$把方程 $y''=q(x)y$ 变换成差分方程

$$y_{k-1} - (2 + h^2 q_k)y_k + y_{k+1} = 0,$$

要求它对 $k=1,\cdots,n$ 相应的自变量$x_1,\cdots,x_n$ 成立.加上 $y_0=A$ 和 $y_{n+1}=B$.我们就有了 n 阶线性方程组.因此得到在列出的自变量处 y 的近似值.

同样,Laplace 方程 $U_{xx}+U_{yy}=0$ 变换成差分方程

$$U(x, y) = \frac{1}{4}[U(x-h, y) + U(x+h, y) + U(x, y-h) + U(x, y+h)],$$

它使得每个值是在正方形网格区域的点 $x_m=x_0+mh$, $y_n=y_0+nh$ 的四个相邻结点上的平均.对每一个内部网格结点写出这个方程就得 N 阶线性方程组.其中 N 是内部网格结点的个数.这种想法适用其他方程,带有曲边边界的区域和高维问题,在相当宽的情形下,能证明对于精确解的收敛性.

经典的扩散方程问题

$$T_t = T_{xx}, \quad T(0, t) = T(1, t) = 0, \quad T(x, 0) = f(x),$$

有限差分方法也适用于它.这个方程在半无限长条区域$0 \leqslant x \leqslant 1, 0 \leqslant t$ 内部成立.在长条的边界上给定 T.它有著名的 Fourier 级数解.但有限差分方法对各种改进很有用处.用简单的差商代替导数,上面的方程就变成

$$T_{m, n+1} = \lambda T_{m-1, n} + (1 - 2\lambda)T_{m, n} + \lambda T_{m+1, n},$$

其中 $x_m=mh, t_n=nk$ 和 $\lambda=k/h^2$.于是矩形网格点集合代替了长条区域,所给形式的差分方程可以从前一时间步的 T 值直接计算每一个 T 值,用所给的初值 $f(x_m)$ 启动这一过程,适当选取 h 和 k,它们趋于零.则方法收敛于真解.但是对于小的 k 计算量很大.因此为了减少工作量已提出了很多改进的方法.

3. **花园浇水法**(garden hose method),对经典的两点边值问题提供了一种直观的近似.我们先解初值问题:

$$y''=f(x,y,y'),\quad y(a)=A,\quad y'(a)=M.$$

M 可以有几种选择,所得到的最终值将依赖于 M 的选择,称它为 $F(M)$,于是我们想要的就是 $F(M)=B$.这一问题类似于第二十五章中的求根问题,因此能用类似的方法解决.求出对 M 的逐次逼近,每一个值都得到一个新的初值问题,如同求根,有好几种方法来选择对 M 的校正,包括牛顿型方法.

$$M_2=M_1-\frac{F(M_1)-B}{F'(M_1)}.$$

4. **变分法**建立某些边值问题和优化问题的等价性.为了求函数 $y(x)$ 满足 $y(a)=A$ 和 $y(b)=B$ 并使

$$\int_a^b F(x,y,y')dx$$

取到最小值(或最大值),我们可以解 Euler 方程

$$\mathrm{F}_y=\frac{d}{dx}\mathrm{F}_{y'},$$

并满足相同的边界条件.还有其他直接方法,例如 Ritz 方法,使积分极小化,因此它可以认为是求解 Euler 方程并且满足边界条件的方法.

对于 Laplace 方程相应的极小化问题是

$$\iint(U_x^2+U_y^2)dxdy=\min,$$

其中二重积分取在边值问题的区域 R 上.

对于 Poisson 方程 $U_{xx}+U_{yy}=K$,相应的优化问题是

$$\iint\left[\frac{1}{2}(U_x^2+U_y^2)+KU\right]\mathrm{d}x\mathrm{d}y=\min.$$

5. **有限元方法**是优化问题直接解法的有效方法.把区域 R 细分成基本小片(对二维的 R 是三角形,正方形等等),解元素和每一个小片区域相结合,例如关于一组基本三角形,我们选取一组平面三角形元素共同形成连续的表面.这些元素的各个角点之纵坐标成为优化的独立变量,使关于这些变量的偏导数等于零,然后再解出所得到的方程组.
6. **无穷级数**为许多经典问题提供了解.它们是叠加原理的发展.重要的有 Fourier 级数及其各种推广.

题　解

线性常微分方程

29.1 求解二阶方程

$$L(y)=y''(x)-p(x)y'(x)-q(x)y(x)=r(x)$$

并满足二个边界条件

$$c_{11}y(a)+c_{12}y(b)+c_{13}y'(a)+c_{14}y'(b)=A,$$
$$c_{21}y(a)+c_{22}y(b)+c_{23}y'(a)+c_{24}y'(b)=B.$$

解 ☞ 对于线性方程,我们可以用叠加原理,这一点已用于解析方法求解的基本例子里,假设对

上述方程不能求出基本解,就能用前几章的数值方法(Runge-Kutta 方法,Adams 方法等等)来计算以下三个初值问题($a\leqslant x\leqslant b$)的近似解

$$L(y_1)=0,\quad L(y_2)=0,\quad L(Y)=r(x),$$
$$y_1(a)=1,\quad y_2(a)=0,\quad Y(a)=0,$$
$$y'_1(a)=0,\quad y'_2(a)=1,\quad Y'(a)=0.$$

于是叠加可得所要求的解是

$$y(x)=C_1y_1(x)+C_2y_2(x)+Y.$$

为了满足边界条件,C_1 和 C_2 可由以下方程组确定:

$$[c_{11}+c_{12}y_1(b)+c_{14}y'_1(b)]C_1+[c_{13}+c_{12}y_2(b)+c_{14}y'_2(b)]C_2$$
$$=A-c_{12}Y(b)-c_{14}Y'(b),$$
$$[c_{21}+c_{22}y_1(b)+c_{24}y'_1(b)]C_1+[c_{23}+c_{22}y_2(b)+c_{24}y'_2(b)]C_2$$
$$=B-c_{22}Y(b)-c_{24}Y'(b).$$

我们关于初值问题的算法以这种方式解出了线性边值问题.这种方法容易推广到高阶方程和线性方程组.我们假定问题有惟一解,而且能求出具有适当精度的 y_1 和 y_2.于是确定 C_1,C_2 等的方程组也有惟一解.

29.2　说明如何通过化为线性代数方程组来近似求解线性边值问题.

解　☞ 选取等距间隔的自变量 $x_j=a+jh$,而且 $x_0=a,x_{N+1}=b$.现在我们企图确定相应的值 $y_j=y(x_j)$,$y''(x_j)$用以下近似代替

$$y''(x_j)\approx\frac{y_{j+1}-2y_j+y_{j-1}}{h^2}.$$

$y'(x_j)$由下式近似

$$y'(x_j)\approx\frac{y_{j+1}-y_{j-1}}{2h}.$$

题 29.1 中的微分方程 $L(y)=r(x)$稍作整理即成为

$$\left(1-\frac{1}{2}hp_j\right)y_{j-1}+(-2+h^2q_j)y_j+\left(1+\frac{1}{2}hp_j\right)y_{j+1}=h^2r_j.$$

如果我们要求它在所有的内点 $j=1,\cdots,N$ 都成立,那么就得到 N 个未知量$y_1,\cdots,y_N$ 的 N 个线性方程,假设两边值规定为 $y_0=y(a)=A,y_{N+1}=y(b)=B$.在这种情形下,线性方程组取以下形式:

$$\beta_1y_1+\gamma_1y_2\qquad=h^2r_1-\alpha_1A,$$
$$\alpha_2y_1+\beta_2y_2+\gamma_2y_3=h^2r_2,$$
$$\alpha_3y_2+\beta_3y_3+\gamma_3y_4=h^2r_3,$$
$$\cdots\cdots\cdots,$$
$$\alpha_Ny_{N-1}+\beta_Ny_n=h^2r_N-\gamma_Nb.$$

其中　$$\alpha_j=1-\frac{1}{2}hp_j,\quad\beta_j=-2+h^2q_j,\quad\gamma_j=1+\frac{1}{2}hp_j.$$

这个方程组的带状矩阵是对微分方程边值问题离散化后得的线性代数方程组的特征.只有三条对角线元素非零.这样的矩阵比其他的不如此稀疏的矩阵要容易处理.若用 Gauss 消去法,因为主元转化成主对角线,所以带的性态不会破坏.可以用这个事实来简化计算.Gauss-Seidel 迭代算法也有效.假定出现题 29.1 中的更一般的边界条件,也能进行处理,多半是用

$$y'(a)\approx\frac{y_1-y_0}{h},\qquad y'(b)=\frac{y_{N+1}-y_N}{h},$$

这就得到了 $N+2$ 个未知量 $y_0,\cdots,y_{N+1}$的 $N+2$ 个方程的方程组.

在这一章以及前面问题中我们已对同一目标进行了可能的逼近.在这两种情形输出的是数 y_j 的一个有限集.如果这两种方法用较小的 h 重复使用.那么所得到的较大的输出有希望更精确地代表真解.这就是收敛性问题.

29.3　证明对特殊情形

$$y''+y=0,\quad y(0)=0,\quad y(1)=1.$$

题 29.2 中的方法是收敛的.

证 ☞ 精确解是 $y(x)=(\sin x)(\sin 1)$. 近似差分方程是

$$y_{j-1}+(-2+h^2)y_j+y_{j+1}=0,$$

对于相同的边界条件 $y_0=0, y_{N+1}=1$ 它的精确解是

$$y_j=\frac{\sin(\alpha x_j/h)}{\sin(\alpha/h)},$$

这里 $x_j=jh, \cos\alpha=1-\frac{1}{2}h^2$. 这些结论可以直接证明或者由本书差分方程这部分的方法推得. 因为当 h 趋于零时, $\lim(\alpha/h)=1$, 所以 $\lim y_j=y(x_j)$. 即当 h 减小时, 差分问题的解收敛于微分方程问题的解. 在这个例子中两个问题可以解析求解. 它们的解可以比较. 对于更一般的问题收敛性的证明必须用其他方法.

29.4 说明线性微分方程特征值问题转化为近似代数方程组

解 ☞ 考虑问题

$$y''+\lambda y=0,\quad y(0)=y(1)=0,$$

它有精确解 $y(x)=C\sin n\pi x, n=1,2,\cdots$. 相应的特征值是 $\lambda_n=n^2\pi^2$. 为简单地说明这种也可用于精确解不易求得的其他问题的方法, 我们用差分方程

$$y_{j-1}+(-2+\lambda h^2)y_j+y_{j-1}=0$$

代替微分方程, 要求它在内点 $j=1,\cdots,N$ 成立. 我们就得到了具有带状矩阵

$$A=\begin{bmatrix}-2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & & 1 \\ & & & 1 & -2\end{bmatrix}$$

的代数特征值问题 $Ay=\lambda h^2 y$. 矩阵 A 的其他元素全为零, $y^T=(y_1,\cdots,y_N)$, 能求出这个问题的精确解是

$$y_j=C\sin n\pi x_j\quad 和\ \lambda_n=\frac{4}{h^2}\sin^2\frac{n\pi h}{2}.$$

显然, 当 h 趋于零时, 这些结果收敛于微分方程特征值问题的结果.

非线性常微分方程

29.5 什么是花园浇水法(garden-hosemethod)?

解 ☞ 对给定的微分方程 $y''=f(x,y,y')$, 我们要求它的满足边界条件 $y(a)=A, y(b)=B$ 的解. 一种简单的方法是计算以下初值问题

$$y''=f(x,y,y'),\quad y(a)=A,\qquad y'(a)=M$$

的解. 对不同的 M 值, 直到求得两个解, 其中一个满足 $y(b)<B$, 而另一个满足 $y(b)>B$. 假如这两个解对应于初始条件 M_1 和 M_2, 那么插入将在这两个值之间提供一个新的值 M. 因此可以算得更好的近似值(见图 29.1). 继续这个过程将不断得到更好的近似值, 这种方法实质上就是用于非线性代数问题的试位算法(regula falsi algorithm). 这里, 我们计算的终值是 M 的函数, 譬如 $F(M)$, 而且, 我们必须解方程 $F(M)=B$. 但是对于 M 的每种选择, $F(M)$ 的算法不再是代数表达式的求值, 而是包括求解一个微分方程初值问题.

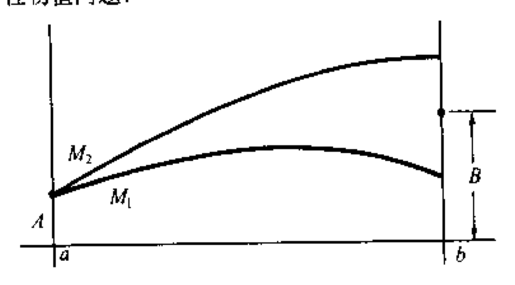


图 29.1

29.6　花园浇水法可以怎样改进?

解　不用相当于试位算法的那种方法，我们对目前的问题可采用牛顿法，希望对正确值 M 的收敛性得到改进. 为此我们要知道 $F'(M)$. 设 $y(x,M)$表示以下问题的解

$$y''=f(x,y,y'),\quad y(a)=A,\quad y'(a)=M.$$

为了简单，令 $z(x,M)$是 $y(x,M)$关于 M 的偏导数. 对 M 求导得到

$$z''=f_y(x,y,y')z+f_{y'}(x,y,y')z',\tag{1}$$

这里我们自由地转换了各个导数的阶数. 再对初始条件求导，得到

$$z(a,M)=0,\quad z'(a,M)=1.$$

设 M_1 是对 M 的首次逼近，然后再求出原问题的近似解 $y(x,M_1)$，于是它可以代替方程(1)中的 y. 而且可以算得 $z(x,M_1)$，因此 $F'(M)=z(b,M_1)$. 利用这个已得到的量解 $F(M)-B=0$ 的牛顿法给出了对 M 的下一个近似：

$$M_2=M_1-\frac{F(M_1)-B}{F'(M_1)}.$$

有了这个 M_2，就可以计算新的近似 $y(x,M_2)$再重复这个过程. 这个方法可推广到高阶方程和方程组. 其核心思想是导出类似(1)的方程，称之为变分方程(variational equation).

优化

29.7　将 $\int_a^b F(x,y,y')\mathrm{d}x$ 的极大或极小问题化为边值问题.

解　这是变分学的经典问题. 如果解函数 $y(x)$存在而且充分光滑，那么它满足 Euler 方程 $F_y=(d/dx)F_{y'}$. 假如在原优化问题中给定边界条件，如 $y(a)=A, y(b)=B$，那么我们就得到了一个二点*边值问题. 如果没有这些条件，那么变分理论指出在区间端点必须成立 $F_{y'}=0$，这称为自然边界条件.

29.8　极小化 $\int_0^1(y^2+y'^2)\mathrm{d}x, y(0)=1$.

解　Euler 方程是 $2y=2y''$，自然边界条件是 $y'(1)=0$. 现在容易求出解是 $y=\mathrm{ch}x-\mathrm{th}1\mathrm{sh}x$ 而且它使该积分等于 th1，(大约等于 0.76). 在一般情形下，Euler 方程是非线性的，可以用花园浇水法求 $y(x)$.

29.9　说明解边值问题的 Ritz 方法.

解　Ritz 方法的思想是解一个等价的极小化问题. 考虑

$$y''=-x^2,\quad y(0)=y(1)=0.$$

它有时被称为单变量的 Poisson 问题. 但是实际上只需要二次积分就得到解

$$y(x)=\frac{x(1-x^3)}{12}.$$

对于给定的边值问题可以得到求与它等价的极小化问题的方法，其中著名的是

$$J(y)=\int_0^1\left[\frac{1}{2}(y')^2-x^2y\right]\mathrm{d}x=\min,$$

对这个积分的 Euler 方程可以证明就是我们原来的微分方程.

为了用 Ritz 方法求近似解，我们需要一族满足边界条件的函数. 假定我们选取

$$\phi(x)=cx(1-x),$$

对这个问题，这可能是这族函数中最简单的. 用 ϕ 代替积分中的 y，简单计算得

$$J(\phi)=\frac{c^2}{6}-\frac{c}{20}=f(c).$$

令 $f'(c)=0$，使其极小化，结果得到 $c=\frac{3}{20}$. 于是我们得到近似解

$$\phi(x)=\frac{3}{20}x(1-x).$$

---

* 译注：原书误为二阶.

在图 29.2 中给出了它与精确解的比较. 还可以通过用更完全的近似函数族得到更精确的近似解. 如果取

$$\phi(x) = x(1-x)(c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n),$$

将导出确定系数 c_i 的线性代数方程组. Ritz 方法的核心思想在限定的函数族 $\phi(x)$ 中寻求优化函数, 而不是在使所给积分存在的所有 $y(x)$ 中寻求优化函数.

29.10 用在题 29.9 中解过的相同的边值问题说明有限元解法.

解 ☞ 基本思想是相同的, 假定我们把区间(0,1)分成两半, 用相交于点 $\left(\frac{1}{2}, A\right)$ 的两条线段

$$\phi_1(x) = 2Ax, \quad \phi_2(x) = 2A(1-x)$$

来近似 $y(x)$. 实际上我们得到了一族这种近似, 其参数 A 可以选择. 这两条线段称为有限元, 而近似解函数是把它们拼在一起形成的. 和前面一样, 我们把它们代入积分得

$$J(\phi) = \int_0^{1/2} \phi_1 dx + \int_{1/2}^1 \phi_2 dx = 2A^2 - \frac{7}{48}A = f(A),$$

可令 $f'(A) = 0$ 使其极小化. 这就得到 $A = \frac{7}{192}$. 计算很快说明这实际上是在 $x = \frac{1}{2}$ 处解的正确值, (见图 29.2). 已经证明如果用这些线段作有限元(当然是在一维问题), 有规则地在交点处得到正确值.

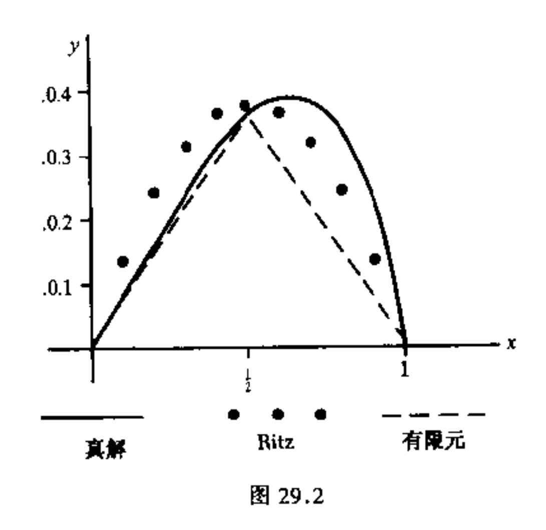


图 29.2

29.11 把上述问题的方法推广到更多的有限元.

解 ☞ 把区间(0,1)分成 n 份, 分点是 $0 = x_0, x_1, x_2, \cdots, x_n = 1$. 设 $y_1, \cdots, y_{n-1}$ 是相应的任意纵坐标. 而且 $y_0 = y_n = 0$. 以明白直观的方式(见图 29.3)定义线性有限元 $\phi_1, \cdots, \phi_n$. 于是

$$\phi_i(x) = y_{i-1}\frac{x_i - x}{x_i - x_{i-1}} + y_i\frac{x - x_{i-1}}{x_i - x_{i-1}} = y_{i-1}\frac{x_i - x}{h} + y_i\frac{x - x_{i-1}}{h}.$$

当 x_i 是等距剖分时, 上面的第二个等式成立. 我们还有

$$\phi_i'(x) = \frac{y_i - y_{i-1}}{x_i - x_{i-1}} = \frac{y_i - y_{i-1}}{h}.$$

现在考虑积分

$$J(\phi) = \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left[\frac{1}{2}(\phi_i')^2 - x^2\phi_i\right]dx = \sum_{i=1}^{n} J_i$$

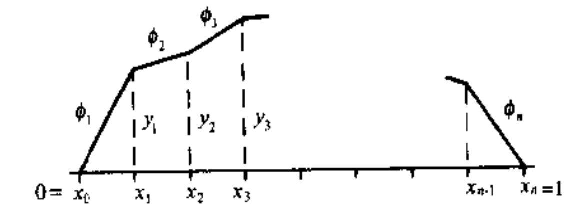


图 29.3

$$= f(y_1, \cdots, y_{n-1}),$$

使其极小化，得到由 y_i 表示的 f，然后再计算它关于 y_i 的偏导数，并令其等于零，再求解所得的方程组. 这就是刚才在较简单情形所做的. 这里假设我们先求导后积分，于是形成了最后的方程组. 对于 f 依赖于一个特定的纵坐标 y_k，仅通过两个部分项 J_k 和 J_{k+1} 与其发生关系. 因此，对 $k=1,\cdots,n-1$，

$$\frac{\partial f}{\partial y_k} = \int_{x_{k-1}}^{x_k}\left[\frac{y_k - y_{k-1}}{h}\left(\frac{1}{h}\right) - x^2\frac{x - x_{k-1}}{h}\right]dx + \int_{x_k}^{x_{k+1}}\left[\frac{y_{k+1} - y_k}{h}\left(\frac{-1}{h}\right) - x^2\frac{x_{k+1} - x}{h}\right]dx.$$

这些积分是初等的，我们立即得到方程组

$$-y_{k-1} + 2y_k - y_{k+1} = \frac{1}{12}x_{k-1}^4 + \frac{1}{2}x_k^4 + \frac{1}{12}x_{k+1}^4 - \frac{1}{3}x_{k-1}x_k^3 - \frac{1}{3}x_{k+1}x_k^3,$$
$$k = 1, \cdots, n-1.$$

当 $n=2, k=1$ 时立即得出前面已得到的 $y=\frac{7}{192}$. 当 $n=3$ 时，方程组是

$$2y_1 - y_2 = \frac{7}{486},$$
$$-y_1 + 2y_2 = \frac{25}{486}.$$

由此可得到 $y_1=\frac{13}{486}, y_2=\frac{19}{486}$. 它们与真解在这两点的是一致的.

扩散方程

29.12　用有限差分近似来代替包括方程

$$\frac{\partial T}{\partial t} = a\left(\frac{\partial^2 T}{\partial x^2}\right) + b\left(\frac{\partial T}{\partial x}\right) + cT$$

和定解条件 $T(0,t)=f(t)$, $T(l,t)=g(t)$, $T(x,0)=F(x)$ 的扩散问题.

解　设 $x_m = mh$, $t_n = nk$，其中 $x_{M+1} = l$. 用符号 $T_{m,n}$ 表示 $T(x,t)$，近似关系式

$$\frac{\partial T}{\partial t} \approx \frac{T_{m,n+1} - T_{m,n}}{k}, \quad \frac{\partial T}{\partial x} \approx \frac{T_{m+1,n} - T_{m-1,n}}{2h},$$
$$\frac{\partial^2 T}{\partial x^2} \approx \frac{T_{m+1,n} - 2T_{m,n} + T_{m-1,n}}{h^2},$$

把扩散方程转变为

$$T_{m,n+1} = \lambda\left(a - \frac{1}{2}bh\right)T_{m-1,n} + [1 - \lambda(2a + ch^2)]T_{m,n} + \lambda\left(a + \frac{1}{2}bh\right)T_{m+1,n},$$

其中 $\lambda = k/h^2$, $m = 1,2,\cdots,M$, $n = 1,2,\cdots$. 利用上面相同的初始和边界条件，记成 $T_{0,n} = f(t_n)$, $T_{M+1,n} = g(t_n)$ 及 $T_{m,0} = F(x_m)$. 这个差分方程给出了每个内点上 $T_{m,n+1}$ 的近似值. 它是用前时间步中它的最近的三个值所表出的. 因此计算从 $t=0$ 所给定的值开始，首先进行到 $t=k$，然后进行到 $t=2k$ 等等.（见下一题的说明.）

29.13　在 $a=1, b=c=0, f(t)=g(t)=0, F(x)=1, l=1$ 的情形下. 用上题的这个方法求解.

解　假设取 $h=\frac{1}{4}, k=\frac{1}{32}$，则 $\lambda=\frac{1}{2}$，于是差分方程化成

$$T_{m,n+1} = \frac{1}{2}(T_{m-1,n} + T_{m+1,n}).$$

表 29.1 (a)摘录了前几行的计算，底行和边列是初始条件和边界条件. 内部值是由部分方程一行接一行计算所得. 从圈圈的①开始，它得自它前一行的两个相邻值，也是圈圈的. 可以注意到慢慢地趋于最终的"稳定状态"，这时所有的 T 值为零. 对如此简单的计算应预期到不太精确.

表 29.1

														0	−12	17	−12	0
														0	5	−7	5	0
0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0	0	$\frac{3}{8}$	$\frac{3}{4}$	$\frac{7}{8}$	1	$\frac{7}{8}$	$\frac{3}{4}$	$\frac{3}{8}$	0	0	−2	3	−2	0
0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$\frac{3}{4}$	1	1	1	$\frac{3}{4}$	$\frac{1}{2}$	0	0	1	−1	1	0
0	$\frac{1}{2}$	1	$\frac{1}{2}$	0	0	$\frac{1}{2}$	1	1	1	1	1	$\frac{1}{2}$	0	0	0	1	0	0
0	①	1	1	0	0	1	1	1	1	1	1	1	0	0	1	1	1	0
①	1	①	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		(a)							(b)							(c)		

其次我们取 $h=\frac{1}{8}$，$k=\frac{1}{128}$保持 $\lambda=\frac{1}{2}$. 结果见表 29.1(b). 这表中最上面一行相当于表 29.1(a)中的第二行，事实上是对 $T\left(x,\frac{1}{32}\right)$ 的更好的近似. 由这个结果得到初步的猜想是这一过程开始收敛于 $T(x,t)$的正确值.

如果我们取 $h=\frac{1}{4}$，$k=\frac{1}{16}$，使 $\lambda=1$. 就有表 29.1(c)中的结果. 这时差分方程是

$$T_{m,n+1} = T_{m-1,n} - T_{m,n} + T_{m+1,n}.$$

可以看到开始激烈地振荡，这并不都和正确解相符合，它称为指数型衰变. 以后，我们将看到除了 $\lambda \leqslant \frac{1}{2}$，都可能出现这类并非实际的振荡. 这是一种数值不稳定的形式.

29.14 这种方法的截断误差是什么？

解 ☞ 和以前把 Taylor 定理用于差分方程一样，发现我们的近似引进了与 h 和 k 有关的误差项. 这些项就是截断误差

$$\frac{1}{2}kT_{tt} - \frac{1}{12}ah^2T_{xxxx} + \frac{1}{6}bh^2T_{xxx} + O(h^4),$$

下标表示偏导数，在重要的特殊情形 $a=1$，$b=0$，我们有 $T_{tt}=T_{xxxx}$，由此看来从使截断误差等于 $O(h^4)$这一观点出发特别要求选取 $k=\frac{h^2}{6}$(或者 $\lambda=\frac{1}{6}$).

29.15 对于特别的情形

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2}, \quad T(0,t) = T(\pi,t) = 0, \quad T(x,0) = \sin px,$$

证明题 29.12 中的方法收敛，其中 p 是正整数.

解 ☞ 可以证明精确解是 $T(x,t)=e^{-p^2t}\sin px$. 相应的差分方程是

$$T_{m,n+1} - T_{m,n} = \lambda(T_{m+1,n} - 2T_{m,n} + T_{m-1,n}).$$

其余的条件可写为

$$T_{m,0} = \sin\frac{mp\pi}{M+1}, \quad T_{0,n} = T_{M+1,n} = 0.$$

能用分离变量法解这个有限差分方程问题. 令 $T_{m,n}=u_mv_n$，代入方程后得到

$$\frac{v_{n+1}-v_n}{v_n} = \lambda\left(\frac{u_{m+1}-2u_m+u_{m-1}}{u_m}\right) = -\lambda C.$$

该方程定义了 C. 把 C 与最左端的项相比，我们发现它与 m 无关，再把 C 与中间部分相比，又发现它与 n 无关，因此它是常数. 我们分别得到 u_m 和 v_n 的方程

$$v_{n+1} = (1-\lambda C)v_n, \quad u_{m+1} - (2-C)u_m + u_{m-1} = 0,$$

它们容易用差分方程方法求解. 除非 $0<C<4$，第二个方程不存在解满足 $u_0=u_{M+1}=0$(除了 $u_m\equiv 0$). 当 $0<C<4$ 时，

$$u_m = A\cos\alpha m + B\sin\alpha m,$$

其中 A 和 B 是常数，$\cos\alpha = 1-\frac{1}{2}C$. 为了满足边界条件，必有 $A=0$ 和 $\alpha(M+1)=j\pi$，j 是整数，因此

$$u_m = B\sin\frac{mj\pi}{M+1}.$$

转向 v_n，我们首先得到 $C=2(1-\cos\alpha)=4\sin^2\{j\pi/[2(M+1)]\}$，然后有

$$v_n=\left[1-4\lambda\sin^2\frac{j\pi}{2(M+1)}\right]^n v_0.$$

现在容易看到取 $B=v_0=1, j=p$ 就得到一个函数

$$T_{m,n}=u_m v_n=\left[1-4\lambda\sin^2\frac{p\pi}{2(M+1)}\right]^n\sin\frac{mp\pi}{M+1},$$

它具有所有要求的性质. 为了和微分方程进行比较，我们回到符号 $x_m=mh, t_n=nk$.

$$T_{m,n}=\left(1-4\lambda\sin^2\frac{ph}{2}\right)^{t_n/\lambda h^2}\sin px_m,$$

假设 $\lambda=k/h^2$ 保持固定，当 h 趋于零时，$\sin px_m$ 的系数有极限 $e^{-p^2t_n}$，因此收敛性得证. 这里我们必须使点 (x_m,t_n) 也保持固定，即当 h 和 k 趋于零时，m 和 n 不断增大，以使得 $T_{m,n}$ 是对同一个 $T(x,t)$ 的逐次近似.

29.16 用上题说明对所考虑的某些特殊情形，除非 $\lambda\leqslant\frac{1}{2}$，可能出现激烈振荡.

解 ☞ 现在的问题不是当 h 趋于零时会发生什么，而是对固定的 h，当计算进行到较大的 n 时会发生什么. 检查 $\sin px_m$ 的系数，我们看到对某些 λ, p 和 h 的值，括号内的量可能比 -1 小，这将导致随着 t_n 的增加而产生激烈的振荡. 要求 $\lambda\leqslant\frac{1}{2}$ 就可以避免这种激烈的振荡. 因为这使 $k\leqslant\frac{h^2}{2}$，所以计算就要进行得很慢. 假定要求对大的 t 的结果，用其他的近似方法可能是有用的(见下一问题).

29.17 用 Fourier 级数解题 29.12.

解 ☞ 当 a 是常数, $b=c=0$ 时，这是一种经典的方法. 我们首先求扩散方程具有乘积形式 $U(x)V(t)$ 的解. 代入方程中得到 $V'/V=U''/U=-\alpha^2$，其中 α 是一常数(负号将有助于满足边界条件)，这样就求得

$$V=Ae^{-\alpha^2t},\qquad U=B\cos\alpha x+C\sin\alpha x.$$

为使 $T(0,t)=0$，我们取 $B=0$. 为使 $T(1,t)=0$，我们取 $\alpha=n\pi$，n 是正整数. 可置 $C=1$ 并且把 A 改写为 A_n，我们得到函数

$$A_ne^{-n^2\pi^2t}\sin n\pi x,\qquad n=1,2,3,\cdots.$$

除了初始条件外，上面的每一个函数都满足我们所有要求，如果级数

$$T(x,t)=\sum_{n=1}^{\infty}A_ne^{-n^2\pi^2t}\sin n\pi x$$

收敛，它当然也满足这些要求，而且通过适当选取 A_n，它也能满足初始条件. 对于 $F(x)=1$ 我们要求

$$T(x,0)=F(x)=\sum_{n=1}^{\infty}A_n\sin n\pi x.$$

这可以用 $F(x)$ 的 Fourier 系数

$$A_n=2\int_0^1F(x)\sin n\pi x\,dx$$

而得到，现在我们的级数的部分和就可以作为扩散方程的近似解. 在题 29.15 中用过的精确解可看成是一项的 Fourier 级数.

Laplace 方程

29.18 用有限差分方程代替 Laplace 方程

$$\frac{\partial^2T}{\partial x^2}+\frac{\partial^2T}{\partial y^2}=0\qquad 0\leqslant x\leqslant l,\quad 0\leqslant y\leqslant l.$$

如果 $T(x,y)$ 的边界值在正方形的四条边上给定，说明怎样得到线性代数方程组.

解 ☞ 自然的近似是

$$\frac{\partial^2 T}{\partial x^2} \approx \frac{T(x-h,y)-2T(x,y)+T(x+h,y)}{h^2},$$

$$\frac{\partial^2 T}{\partial y^2} \approx \frac{T(x,y-h)-2T(x,y)+T(x,y+h)}{h^2}.$$

它们立即导出差分方程

$$T(x,y) = \frac{1}{4}[T(x-h,y)+T(x+h,y)+T(x,y-h)+T(x,y+h)],$$

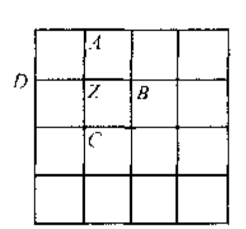


图 29.4

它要求每一个 T 值是它的四个相邻点上的值的平均值.这里我们把注意力集中到水平和垂直距离都是 h 的正方形网格.我们的差分方程就能简化为

$$T_Z = \frac{1}{4}(T_A + T_B + T_C + T_D),$$

只有在图 29.4 中所示出的点.对每一个内点 Z,写出这样的方程(T 是未知量),我们就得到了每一个方程包含五个未知量(除了当已知的边界值减少了这个数目)的线性代数方程组.

29.19　当 $T(x,0)=1$,其余边界值为零时,应用上题中的方法.

解　☞ 为简单计,我们选取 h,使得只有九个内点,如图 29.4,从顶行开始.从左到右给出这些内点排序,我们的九个方程如下:

$$T_1 = \frac{1}{4}(0+T_2+T_4+0),\qquad T_6 = \frac{1}{4}(T_3+0+T_9+T_5),$$

$$T_2 = \frac{1}{4}(0+T_3+T_5+T_1),\qquad T_7 = \frac{1}{4}(T_4+T_8+1+0),$$

$$T_3 = \frac{1}{4}(0+0+T_6+T_2),\qquad T_8 = \frac{1}{4}(T_5+T_9+1+T_7),$$

$$T_4 = \frac{1}{4}(T_1+T_5+T_7+0),\qquad T_9 = \frac{1}{4}(T_6+0+1+T_8).$$

$$T_5 = \frac{1}{4}(T_2+T_6+T_8+T_4),$$

这个方程组可以用 Gauss 消去法求解,但是用 Gauss-Seidel 迭代更自然.从每一个内点很差的初始近似 0 开始得到了在表 29.2 中给出的逐次得到的结果.对这个方程组迭代十次就得到三位精度.(关于 Gauss Seidel 迭代收敛性的讨论见题 26.34.)

表 29.2

迭代次数	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0.250	0.312	0.328
2	0	0	0	0.062	0.078	0.082	0.328	0.394	0.328
3	0.016	0.024	0.027	0.106	0.152	0.127	0.375	0.464	0.398
4	0.032	0.053	0.045	0.140	0.196	0.160	0.401	0.499	0.415
5	0.048	0.072	0.058	0.161	0.223	0.174	0.415	0.513	0.422
6	0.058	0.085	0.065	0.174	0.236	0.181	0.422	0.520	0.425
7	0.065	0.092	0.068	0.181	0.244	0.184	0.425	0.524	0.427
8	0.068	0.095	0.070	0.184	0.247	0.186	0.427	0.525	0.428
9	0.070	0.097	0.071	0.186	0.249	0.187	0.428	0.526	0.428
10	0.071	0.098	0.071	0.187	0.250	0.187	0.428	0.526	0.428

收敛性证明

29.20　证明在题 29.18 中遇到的线性代数方程组存在惟一解.

证 ☞ 关键是因为我们的近似解是基于这个方程组，所以它的非奇性很重要. 把内点上的未知值记为 $T_1,\cdots,T_N$，我们可以把方程组写为以下形式

$$\sum_{k=1}^{N} a_{ik}T_k = b_i, \tag{1}$$

其中 b_i 与边界值有关. 假设边界值都是零，那么 b_i 也将都是零：

$$\sum_{k=1}^{N} a_{ik}T_k = 0. \tag{2}$$

根据线性代数的基本定理，假定(2)仅有零解，那么方程组(1)就有惟一解. 如果极大值 T_k 出现在内点 Z，于是，因为 $T_Z=\frac{1}{4}(T_A+T_B+T_C+T_D)$，那么它也必然出现在 Z 的邻点 A,B,C,D. 类似地，这个极大值也将出现在 A,B,C,D 本身的邻点. 继续这种讨论，我们发现 T_k 的极大值必然出现在边界. 因此必为零，同样的讨论可证明 T_k 的极小值必然出现在边界，因此必为零. 因此方程组(2)中的 T_k 都是零，于是基本定理适用. 注意，我们的证明包含了一个附带的定理，即对(1)和(2)的 T_k 的最大最小值都出现在边界点上.

29.21 证明当 h 趋于零时，题 29.20 中的方程组(1)的解收敛于相应的 Laplace 方程的解.

证 ☞ 用 $T(x,y,h)$ 表示(1)的解，用 $T(x,y)$ 表示 Laplace 方程的解，两者的边界条件相等. 当 h 趋于零时，我们证明在每一点 (x,y) 成立

$$\lim T(x,y,h) = T(x,y).$$

为了方便我们引入符号

$$L[F]=F(x+h,y)+F(x-h,y)+F(x,y+h)$$
$$+F(x,y-h)-4F(x,y),$$

在上式右端用 Taylor 定理，我们容易发现当 $F=T(x,y)$ 时，$|L[T(x,y)]|\leqslant Mh^4/6$，这里 M 是 $|T_{xxxx}|$ 和 $|T_{yyyy}|$ 的上界. 而且，根据定义知 $L[T(x,y,h)]=0$. 现在假设 $x-y$ 坐标系的原点是我们正方形的左下角顶点. 这可以通过坐标变换而得到安排，这样并不改变 Laplace 方程. 引入函数

$$S(x,y,h)=T(x,y,h)-T(x,y)-\frac{\Delta}{2D^2}(D^2-x^2-y^2)-\frac{\Delta}{2},$$

其中 Δ 是任一正数，D 是正方形对角线的长，直接计算可得到

$$L[S(x,y,h)]=\frac{2h^2\Delta}{D^2}+O\left(\frac{Mh^4}{6}\right),$$

因此对充分小的 h，$L[S]>0$. 这表示 S 不能在正方形的内点取到其极大值. 因此极大值在边界上取到. 但是在边界上 $T(x,y,h)=T(x,y)$，而且 S 一定是负的. 这样就使 S 在任何地方都是负的. 因此，我们容易推得 $T(x,y,h)-T(x,y)<\Delta$. 对函数

$$R(x,y,h)=T(x,y)-T(x,y,h)-\frac{\Delta}{2D^2}(D^2-x^2-y^2)-\frac{\Delta}{2}$$

进行类似的讨论，可以证明 $T(x,y)-T(x,y,h)<\Delta$. 这两个结果合起来就表示当 h 充分小时，$|T(x,y,h)-T(x,y)|<\Delta$ 对任意小的 Δ 成立. 这就是收敛性.

29.22 证明如在题 29.19 中应用 Gauss-Seidel 方法，收敛于题 29.20 中方程组(1)的精确解 $T(x,y,h)$.

证 ☞ 当然这与刚才得到的收敛性结果是完全不同的. 这里我们关心的是 $T(x,y,h)$ 的实际计算. 而且已经选好一种逐次求近似解的方法. 假定我们把正方形网格的内点从 1 到 N 排序如下：首先把在顶排的点从左排到右，然后对下一排的点从左排到右，等等. 在所有的内点给定任意的初值 $T_i^0(i=1,\cdots,N)$，后面的近似值称为 T_i^n，当 n 趋于无穷时，我们要证明

$$\lim T_i^n = T_i = T(x,y,h).$$

设 $S_i^n=T_i^n-T_i$. 现在的目的是证明 $\lim S_i^n=0$. 证明是根据这样的事实，即每个 S_i 是它的四个相邻点上值的平均，这是因为 T_i^n 和 T_i 都有这个性质.（在边界点我们置 S 等于零.）设 M 是 $|S_i^0|$ 的最大值. 因为第一个点至少与一个边界点相邻，所以

$$|S_1^1|\leqslant\frac{1}{4}(M+M+M+0)=\frac{3}{4}M,$$

而且每一个后面的点至少与前一点相邻，所以

$$|S'_{i+1}| \leqslant \frac{1}{4}(M + M + M + |S'_i|)$$

为了用归纳法，可设$|S'_i| \leqslant \left[1 - \left(\frac{1}{4}\right)^i\right]M$，我们立即就有

$$|S'_{i+1}| \leqslant \frac{3}{4}M + \frac{1}{4}\left[1 - \left(\frac{1}{4}\right)^i\right]M = \left[1 - \left(\frac{1}{4}\right)^{i+1}\right]M,$$

这样就完成了归纳法．即有$|S'_N| \leqslant \left[1 - \left(\frac{1}{4}\right)^N\right]M = \alpha M$，它进一步表明

$$|S'_i| \leqslant \alpha M \quad i = 1, \cdots, N.$$

重复这一过程就得到$|S''_i| \leqslant \alpha^n M$．因为$\alpha < 1$，就有$\lim S''_i = 0$．这正是所需要的．虽然这里是对任意初值$T^0_i$证明了收敛性，但是，如果我们找到精确的初始值，显然能更迅速地得到T^n_i更好的近似

29.23 对 Poisson 方程

$$U_{xx} + U_{yy} = K \quad (K \text{ 是常数})$$

用三角形元建立有限元的基本公式．

解 ☞ 必须先把方程成立的区域分成三角形小块，再在所需要处进行近似．令(x_i, y_i)，(x_j, y_j)，(x_k, y_k)是这种三角形的顶点，在这个三角形上的解曲面被平面元素$\phi^{(e)}(x, y)$所近似．它的上标和问题中的元素有关．假设z_i, z_j, z_k是三角形顶点到这平面的距离，于是

$$\varphi^{(e)} = L_i^{(e)}z_i + L_j^{(e)}z_j + L_k^{(e)}z_k,$$

其中$L_i^{(e)}$在结点i处等于1，而在其余两个结点处它等于零，$L_j^{(e)}$和$L_k^{(e)}$具有相应的性质．设Δ_e是由这三个结点组成的基本三角形的面积．因此，

$$2\Delta_e = \begin{vmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix},$$

它很快可得出以下表达式

$$L_i^{(e)} = \frac{1}{2\Delta_e}\begin{vmatrix} 1 & x & y \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix}, \quad L_j^{(e)} = \frac{1}{2\Delta_e}\begin{vmatrix} 1 & x & y \\ 1 & x_k & y_k \\ 1 & x_i & y_i \end{vmatrix}, \quad L_k^{(e)} = \frac{1}{2\Delta_e}\begin{vmatrix} 1 & x & y \\ 1 & x_i & y_i \\ 1 & x_j & y_j \end{vmatrix}.$$

如果写成

$$L_i^{(e)} = \frac{1}{2\Delta_e}(a_i + b_i x + c_i y),$$

于是从行列可得到

$$a_i = x_j y_k - x_k y_j, \quad b_i = y_j - y_k, \quad c_i = x_k - x_j.$$

从$L_j^{(e)}$和$L_k^{(e)}$可得到与上面相平行的公式

$$a_j = x_k y_i - x_i y_k, \quad b_j = y_k - y_i, \quad c_j = x_i - x_k,$$

$$a_k = x_i y_j - x_j y_i, \quad b_k = y_i - y_j, \quad c_k = x_j - x_i.$$

所有的系数a, b, c应该都有上标(e)，但是为简单起见已把它省掉了．

现在考虑和 Poisson 方程等价的极小化问题

$$J(U) = \iint \left[\frac{1}{2}(U_x^2 + U_y^2) + KU\right]\mathrm{d}x\mathrm{d}y = \min,$$

它在边值问题的给定区域上求二重积分．我们用函数ϕ来近似U，ϕ是各自定义于R的三角形剖分的平面三角形元素的组合，所以我们考虑以下极小化问题的替代问题

$$J(\phi) = \sum J_e(\phi^{(e)}).$$

和的每一项在它本身的基本三角形求值．我们想置$J(\phi)$的适当的导数等于零．为此目的，要求J_e分量的导数．注意到

$$\phi_x^{(e)} = \frac{1}{2\Delta_e}(b_i z_i + b_j z_j + b_k z_k),$$

$$\phi_y^{(e)} = \frac{1}{2\Delta_e}(c_i z_i + c_j z_j + c_k z_k),$$

所以省略上标就得到

$$J_e=\iint\left[\frac{1}{2}(\phi_x^2+\phi_y^2)+K\phi\right]\mathrm{d}x\mathrm{d}y=f(z_i,z_j,z_k),$$

求导数是直接了当的,例如

$$\frac{\partial f}{z_i}=\iint\left\{\phi_x\frac{b_i}{2\Delta_e}+\phi_y\frac{c_i}{2\Delta_e}+KL_i\right\}\mathrm{d}x\mathrm{d}y$$

$$=\frac{1}{\Delta_e}\left(\frac{b_i^2+c_i^2}{4}z_i+\frac{b_ib_j+c_ic_j}{4}z_j+\frac{b_ib_k+c_ic_k}{4}z_k\right)+\frac{4}{3}\Delta_e.$$

对$\frac{\partial f}{z_j}$和$\frac{\partial f}{z_k}$有非常类似的结果.这三个结果可简洁地以矩阵形式结合起来

$$\begin{bmatrix}\partial f/z_i\\ \partial f/z_j\\ \partial f/z_k\end{bmatrix}=\frac{1}{4\Delta_e}\begin{bmatrix}b_i^2+c_i^2 & b_ib_j+c_ic_j & b_ib_k+c_ic_k\\ b_ib_j+c_ic_j & b_j^2+c_j^2 & b_jb_k+c_jc_k\\ b_ib_k+c_ic_k & b_jb_k+c_jc_k & b_k^2+c_k^2\end{bmatrix}\begin{bmatrix}z_i\\ z_j\\ z_k\end{bmatrix}+\frac{4}{3}\Delta_e\begin{bmatrix}1\\1\\1\end{bmatrix}.$$

假设 K 是常数这个事实使所需的积分很容易得到这个结果.根据初等微积分也可知每一函数 L 的积分是 1/3.

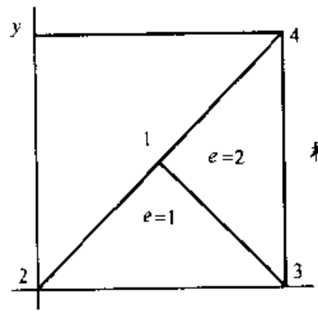
上面的矩阵方程包括把 J 的偏导数集合起来的需要成分.在个别应用时,仍要进行正确的集合.特别地对每一元素 $\phi^{(e)}$,必须注意起作用的点 i,j,k.而且记录关于相应变量 $z_1,z_2,x_3,\cdots$的导数的贡献.

29.24 对上题用有限元法,给定的区域 R 是图 29.5 中的单位正方形,边界值给定,容易看到精确解是 $U(x,y)=x^2+y^2$,它满足 $U_{xx}+U_{yy}=4$.

解 ☞ 因对称性,只需考虑正方形的右下半,而且这已分成两个三角形,结点编号从 1 到 4.这两个三角形由所含的结点编号数所确定.

结点	x	y
1	$\frac{1}{2}$	$\frac{1}{2}$
2	0	0
3	1	0
4	1	1

元素(由结点编号数)
1　2　3　($e=1$)
1　3　4　($e=2$)
$\Delta_1=\Delta_2=\frac{1}{4}$



相应于 $U(x,y)=x^2+y^2$ 的边界值

图 29.5

根据这种基本输入信息,我们先计算系数 a,b,c.下面每列对应于一个结点(i,j,k)

	$e=1$			$e=2$		
a	0	$\frac{1}{2}$	0	1	0	$-\frac{1}{2}$
b	0	$-\frac{1}{2}$	$\frac{1}{2}$	-1	$\frac{1}{2}$	$\frac{1}{2}$
c	1	$-\frac{1}{2}$	$-\frac{1}{2}$	0	$-\frac{1}{2}$	$\frac{1}{2}$

证明每一列的确提供了所要的函数 $L_i^{(e)}$ 是有用的.例如第一列给出

$$L_1^{(1)} = 2[0 - (0)x + (1)y],$$

其中首项2是因为 $1/2\Delta_e$，在结点1，它的值是1，而在结点2和3它是零.用同样的方法可以证明其他列.

为清楚起见，现在将比可能需要更详细地提出 $J(\phi) = f(z_1, z_2, z_3, z_4)$ 的偏导数的集合过程.前一题的矩阵方程包括了来自两个基本元中的每一个的这些导数的贡献.由元素1得到

	z_1	z_2	z_3	
$\partial f/z_1$	1	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{3}$
$\partial f/z_2$	$-\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{3}$
$\partial f/z_3$	$-\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{3}$

最后一列包括常数.元素2得到

	z_1	z_3	z_4	
$\partial f/z_1$	1	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{3}$
$\partial f/z_2$	$-\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{3}$
$\partial f/z_3$	$-\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{3}$

把这两个矩阵合起来就得到

	z_1	z_2	z_3	z_4	
$\partial f/z_1$	2	$-\frac{1}{2}$	-1	$-\frac{1}{2}$	$\frac{2}{3}$
$\partial f/z_2$	$-\frac{1}{2}$	$\frac{1}{2}$	0	0	$\frac{1}{3}$
$\partial f/z_3$	-1	0	1	0	$\frac{2}{3}$
$\partial f/z_4$	$-\frac{1}{2}$	0	0	$\frac{1}{2}$	$\frac{1}{3}$

有了这种对集合元素的说明，现在必须承认，对目前情形，仅仅顶行是实际需要的. z_2, z_3, z_4 的值是边界值，给定为0，1，2.它们不是独立变量，函数 f 仅依赖于 z_1.置这一导数为0，并插入边界值，就得到

$$2z_1 - \frac{1}{2}(0) - (1) - \frac{1}{2}(2) + \frac{2}{3} = 0,$$

使 $z_1 = \frac{2}{3}$.当然准确值是 $\frac{1}{2}$.

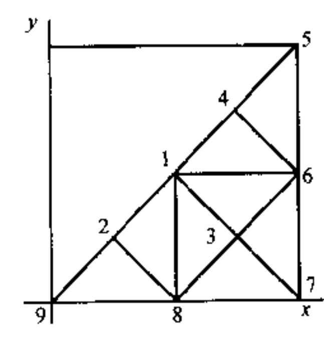
29.25　用在图29.6中给出的更好的三角形网格，重新做前题.

解　我们有这些输入部分：首先，结点1到4，它们的坐标是 $\left(\frac{1}{2}, \frac{1}{2}\right)$，$\left(\frac{1}{4}, \frac{1}{4}\right)$，$\left(\frac{3}{4}, \frac{1}{4}\right)$ 和 $\left(\frac{3}{4}, \frac{3}{4}\right)$，相应的坐标 z 要确定；第二，结点5到9，在这些点处赋于边界值使得 (x, y, z) 的坐标是 $(1,1,2)$，$\left(1, \frac{1}{2}, \frac{5}{4}\right)$，$(1,0,1)$，$\left(\frac{1}{2}, 0, \frac{1}{4}\right)$ 和 $(0,0,0)$；第三，由结点编号数确定八个基本三角形：

298, 281, 183, 387, 376, 136, 164, 465.

如上所描述的有限元算法的计算程序将需要这种输入信息.

假设我们开始手工计算,首先仅在八个元素中的一个进行计算,系数 a, b, c 证明是如下:



a	0	$\frac{1}{8}$	0
b	0	$-\frac{1}{4}$	$\frac{1}{4}$
c	$\frac{1}{2}$	$-\frac{1}{4}$	$-\frac{1}{4}$

图 29.6

这可以像在前题中一样进行检查,各列代表按给定次序的三个结点.每一个基本三角形的面积是 $\frac{1}{16}$.因为所需要的偏导数仅与 z_1 到 z_4 有关.所以我们可以通过寻找对这些导数有贡献的项来减少手工作业.对于这种元素我们有

$$b_i^2 + c_i^2 = 0 + \frac{1}{4} = \frac{1}{4},$$

$$b_ib_j + c_ic_j = 0 - \frac{1}{8} = -\frac{1}{8}, \qquad b_ib_k + c_ic_k = 0 - \frac{1}{8} = -\frac{1}{8}.$$

上面在乘以 $1/4\Delta_e = 4$ 后,进入偏导数矩阵的第二、第八和第九列.常数 $4\Delta_e/3 = \frac{1}{12}$ 也被记录下来.所有第二行中的表值都与 $\frac{\partial f}{\partial z_2}$ 有关

	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	
$\partial f/z_1$	$\frac{1}{2}$	$-\frac{1}{2}$						0		
$\partial f/z_2$		1						$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{12}$
$\partial f/z_3$										
$\partial f/z_4$										

还要找出其他七个元素的类似贡献.并把它们组合到上述矩阵中去,验证第二个元素引入在第一行中的项并找出它对第二行进一步的贡献是有用的.其余的组合过程将留给计算机,就像边界值的代换和所得到四阶线性方程组的求解.我们得到以下输出数据:

结点	计算解	真解
1	0.500000	$\frac{1}{2}$
2	0.166667	$\frac{1}{8}$
3	0.666667	$\frac{5}{8}$
4	1.166667	$\frac{9}{8}$

在第一个结点处的击中靶心完全吻合(牛眼:bull's-eye)是有趣的.

29.26 对四分之一圆域内的 Poisson 方程用同样的有限元方法求解,采用在图 29.7 显示的单个元素.

和边界值 $x^2+y^2=1$ 样，又一次用到 Poisson 方程. 因此真解和 x^2+y^2 相同.

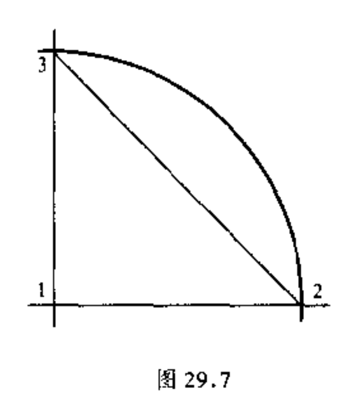


图 29.7

解 ☞ 这一问题说明可以用直线来近似表示曲边边界. 一般地将要用很多这种线段. 三个结点的坐标是:

结点	x	y	z
1	0	0	—
2	1	0	1
3	0	1	1

z_1 的值是优化过的独立变量. 系数 a,b,c 是

	结点 1	结点 2	结点 3
a	1	0	0
b	−1	1	0
c	−1	0	1

由此导出

$$\frac{\partial f}{z_1}=z_1-\frac{1}{2}z_2-\frac{1}{2}z_3+\frac{2}{3}.$$

由此立即可得 $z_1=\frac{1}{3}$. 当然正确值是零，根据对称性，用四个这种三角形就能对整个圆求得相同的结果.

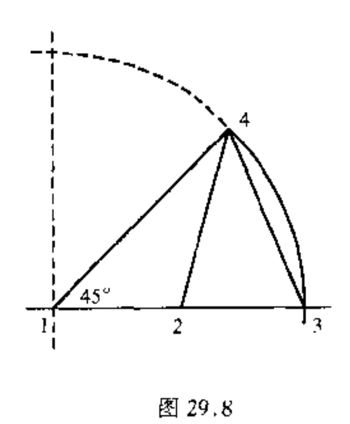


图 29.8

图 29.9

29.27 当使用有限元方法时，通过比较由图 29.8 和图 29.9 中所表出的两个三角形和四个三角形作为元素所求得的粗近似. 叙述收敛性的概念.

解 ☞ 不用说，所有这些结果都比较粗，但观察其结果是有趣的.

结点	(0,0)	$\left(\frac{1}{2},0\right)$
图 29.7	0.33	–
图 29.8	–0.08	0.35
图 29.9	–0.03	0.26
真解	0	0.25

事情已开始向好的方向发展. 已经证明只要合理地改进元素，有限元方法是收敛的.

波动方程

29.28 把有限元法用于方程

$$\frac{\partial^2 U}{\partial t^2}-\frac{\partial^2 U}{\partial x^2}=F[t,x,U,U_t,U_x],\quad -\infty<x<\infty,0\leqslant t,$$

其初始条件是 $U(x,0)=f(x)$， $U_t(x,0)=g(x)$.

解 ☞ 引入矩形网格结点 $x_m=mh$，$t_n=nk$. 在 $t_n=0$ 处 U 的值由初始条件给定. 利用

$$\frac{\partial U}{\partial t}\approx\frac{U(x,t+k)-U(x,t)}{k},$$

在 $t=0$ 处我们有 $U(x,k)\approx f(x)+kg(x)$. 进行到较高的 t 层时，我们需要用微分方程，多半用以下的近似

$$\frac{U(x,t+k)-2U(x,t)+U(x,t-k)}{k^2}$$
$$-\frac{U(x+h,t)-2U(x,t)+U(x-h,t)}{h^2}$$
$$=F\left[t,x,U,\frac{U(x,t)-U(x,t-k)}{k},\frac{u(x+h,t)-U(x-h,t)}{2h}\right].$$

由它可以解出 $U(x,t+k)$. 逐次对 $t=k,k+1,\cdots$，应用这个公式，就产生了 U 在 t 层上所有 x_m 上的 U 值.

29.29 对于简单情形 $F=0,f(x)=x^2,g(x)=1$ 叙述上面的方法.

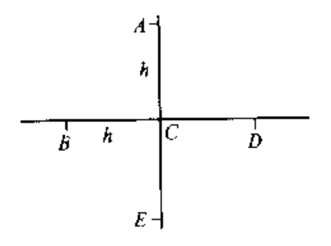


图 29.10

解 ☞ 基本差分方程可写为(见图 29.10)

$$U_A=2(1-\lambda^2)U_C+\lambda^2(U_B+U_D)-U_E,$$

其中 $\lambda=k/h$. 对于 $\lambda=1$，这个差分方程特别简单. 在表 29.3 中给出了 $h=k=0.2$ 时的计算结果. 注意 $x=0$ 到 $x=1$ 的初值确定了一个近似三角形区域的 U 值. 这点对微分方程也是对的，$U(x,t)$的值由位于$(x-t,0)$和$(x+t,0)$之间的初值所决定.(见题 29.30.)

表 29.3

0.6			1.00	1.20		
0.4		0.52	0.64	0.84	1.12	
0.2	0.20	0.24	0.36	0.56	0.84	1.20
0	0.00	0.04	0.16	0.36	0.64	1.00
t/x	0	0.2	0.4	0.6	0.8	1.0

29.30 证明：具有 $U_{tt}=U_{xx}$，$U(x,0)=f(x)$，$U_t(x,0)=g(x)$的精确解 $U(x,t)$的值，依

赖于$(x-t,0)$和$(x+t,0)$之间的初始值.

证 ☞ 对这个古老的熟悉问题,这里是作为试验情形,容易证明精确解是

$$U(x,t)=\frac{f(x+t)+f(x-t)}{2}+\frac{1}{2}\int_{x-t}^{x+t}g(\xi)\mathrm{d}\xi,$$

因此就立即得到所需结论.对更一般的问题成立类似的结论.

29.31 对目前的例子叙述收敛性概念.(注:应是对题29.29中的情况.)

解 ☞ 保持$\lambda=1$,我们减小步长h和k.作为开始,当$h=k=0.1$时的一些结果列在表29.4中,圈起来的数是对$U(0.2,0.2)$的第二个近似值,因此0.26大概是比0.24更精确.用$h=k=0.05$对这个位置的U将得到值0.27.因为微分方程的精确解可以证明是

$$U(x,t)=x^2+t^2+t.$$

我们看到$U(0.2,0.2)=0.28$,而且当h和k减小时,我们的计算是朝着正确解的方向进行的.但是这决不是证明收敛性.同样,另一个圈圈的是对$U(0.4,0.4)$的第二个近似.它比我们早先的0.64要好,因为正确值是0.72.

表 29.4

0.4				0.61	0.68			
0.3			0.40	0.45	0.52	0.61		
0.2		0.23	0.26	0.31	0.38	0.47	0.58	
0.1	0.10	0.11	0.14	0.19	0.26	0.35	0.46	0.59
0	0.00	0.01	0.04	0.09	0.16	0.25	0.36	0.49
t/x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7

29.32 尽管选取$\lambda=k/h>1$可以使在t方向进行得更快,为什么我们不介绍它?

解 ☞ 正确解$U(x,t)$的值依赖于位于$(x-t,0)$和$(x+t,0)$之间的初值,如果$\lambda>1$,计算在(x,t)处的值将仅依赖于这区间的子集AB上的初值(见图29.11),可以改变AB之外的初值,这将影响到其解.但是不影响我们在(x,t)处计算得到的值.这和事实不符.

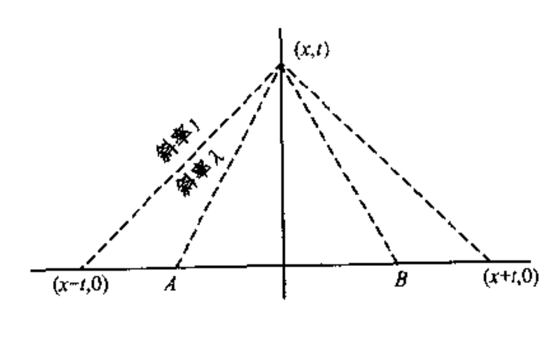


图 29.11

补 充 题

29.33 用题29.1的方法求解方程$y''+y'+xy=0$及$y(0)=1$和$y(1)=0$.

29.34 用题29.2的方法求解前题.你觉得哪种方法更方便?

29.35 解$y''+\sqrt{x}y'+y=e^x$, $y(0)=0$和$y(1)=0$.

29.36 对$y''+\lambda y=0$及$y(0)=0$, $y'(1)=0$,用题29.4中的方法证明对正确解$y=\sin(2n+1)(\pi x/2)$, $\lambda_n=[(2n+1)(\pi/2)]^2$的收敛性.

29.37 用题29.4的方法求$y''+\lambda xy=0$, $y(0)=y(1)=0$的最大特征值.

29.38 对$y''=y^2+(y')^2$, $y(0)=0$, $y(1)=1$用题29.5的方法.

29.39 一物体在一秒内从地平线上升至100英尺高,假定大气阻力的影响,使得运动方程是 $y''=-32-0.1\sqrt{y'}$,求初始速度.

29.40 一物体在一秒内从(0,0)上升到(2000,1000),长度为英尺,假定运动方程是

$$x''(t)=-1.\sqrt{v}\cos\alpha,\quad y''(t)=-32-0.1\sqrt{v}\sin\alpha,$$

其中 $v^2=(x')^2+(y')^2$, $\alpha=\arctan(y'/x')$,求初始速度.

29.41 用题29.7的方法求函数 $y(x)$,使得 $\int_0^1[xy^2+(y')^2]\mathrm{d}x$ 取极小,并且满足 $y(0)=0, y(1)=1$.

29.42 把题29.12中的方法用到情形 $a=c=1, b=0, l=1, f(t)=g(t)=0, F(x)=x(1-x)$.取 $\lambda=\frac{1}{2}$.不断减小 h,逐次得到近似解,直到你觉得已经精确到两位小数.

29.43 取 $\lambda=\frac{1}{6}$,重复前一题,得到满意的结果是更经济还是并不经济?再取 $\lambda=1$ 试验一次.

29.44 证明通过简单的有限差商代替导数把二维扩散方程 $T_t=T_{xx}+T_{yy}$ 变换为

$$T_{l,m,n+1}=(1-4\lambda)T_{l,m,n}+\lambda(T_{l+1,m,n}+T_{l-1,m,n}+T_{l,m+1,n}+T_{l,m-1,n}),$$

并且对三维扩散方程 $T_t=T_{xx}+T_{yy}+T_{zz}$ 得到类似的近似.

29.45 对于在区域 $0\leqslant x, 0\leqslant y, y\leqslant 1-x^2$ 中的Laplace方程及 $T(0,y)=1-y, T(x,0)=1-x$,其余的边界值为零,求近似解.处理曲边边界用最简单的方法,即把边界值转移到附近的网格点.取 $h=\frac{1}{4}$ 和 $h=\frac{1}{8}$ 作试验.你认为你的结果的精确程度如何?

29.46 采用Ritz近似 $\phi(x)=x(1-x)(c_0+c_1x)$ 重复题29.9中的办法.画出相应的曲线并与真解进行比较.

29.47 对 $n=4$ 的情形,写出题29.11中的线性方程组.把它解出来并证明求得了精确值.

29.48 证明关于 z_i, z_j, z_k 的偏导数和题29.23中给出的相同.

29.49 证明系数 a,b,c 和在题29.24中提出的是一样的.

29.50 证明第二个有限元的贡献和在题29.25中提出的是一样的.

29.51 证明题29.27中对两个三角形和四个三角形所给出的结果.

29.52 把有限元方法用于Laplace方程(置 $K=0$ 而不是4)在以(0,0),(1,1),(-1,1)为顶点的三角形区域,边界值由 y^2-x^2 给出.注意,这使得 $U(x,y)=y^2-x^2$ 是真解.根据对称性知只须在三角形的右半部计算就足够了.用两个内部结点 $\left(0,\frac{1}{3}\right)$, $\left(0,\frac{2}{3}\right)$ 和(1,1)一起形成三个基三角形,U 的正确值在这两个内点当然是 $\frac{1}{9}$ 和 $\frac{4}{9}$.这三个元素产生什么值呢?

29.53 对 $T_{xx}+T_{yy}+T_{zz}=0$ 提出一种简单的有限差分逼近.

29.54 边值问题 $y''=n(n-1)y/(x-1)^2, y(0)=1, y(1)=0$ 有一个基本解,不顾这一事实,用花园浇水法求解.取 $n=2$.

29.55 对 $n=20$ 再试一次上题,麻烦的性质是什么?

29.56 边值问题 $y''-n^2y=-n^2/(1-e^{-n}), y(0)=0, y(1)=1$ 有一个基本解.不顾这一事实,用一种我们的近似方法求解,取 $n=1$.

29.57 取 $n=100$ 试验前题.麻烦是什么?

29.58 边值问题

$$U_{tt}+U_{xxxx}=0,\quad 0<x\quad 0<t,$$

$$U(x,0)=U_t(x,0)=U_{xx}(0,t)=0, U(0,t)=1$$

表示木梁振动,起初在 x 轴上静止,且在 $x=0$ 处给了一个位移,这个问题能用Laplace变换求解,其结果表为Fresnel积分,这个积分必须用数值积分来计算.现在用一种我们的有限差分方法来求解.

第三十章 Monte Carlo 方法

随机数

对我们的目的而言，随机数不是由随机过程(如硬币的投掷或轮子的旋转)生成的数.相反它们是由完全确定的计算过程所产生的，结果所得到数集有各种统计性质，它们统称为随机性.典型的过程是

$$x_{n+1} = rx_n(\mathrm{mod}N),$$

其中 r 和 N 给定，x_0 是"随机"数序列的"开端".这种模乘积方法常用作随机数的生成程序.对十进制计算机采用

$$x_{n+1} = 7^9 x_n(\mathrm{mod}\ 10^s), \quad x_0 = 1.$$

对二进制计算机，当 t 是大数时采用

$$x_{n+1} = (8t - 3)x_n(\mathrm{mod}\ 2^s), \quad x_0 = 1.$$

某些生成程序包括以下形式的附加成分

$$x_{n+1} = (rx_n + s)(\mathrm{mod}\ N).$$

对实际问题合适的简单例子是

$$x_{n+1} = (25,173x_n + 13,849)(\mathrm{mod}\ 65,536),$$

由它产生从 0 到 65,535 的整数的 Well-scrambled 分布.

数字 x_n 的序列要被认为是随机数必须通过一组统计试验，它们必须均匀地分布在区间 $(0, N)$，必须有预期的上下两束数字(例如 13,69,97)或者三束数字(09,17,21,73)等等.有时合格的序列据说是由伪随机数组成.大概是把随机这个词留给真正的随机装置(轮盘赌)的产品，在这一章中随机性将表示产品的性质而不是生成器的性质.这将在术语上掩盖住表面的矛盾.许多程序语言，(例如 Fortran)都有装在内部的可调用的随机数生成器，很像它被造成一个模数乘法装置.

应用

通过使用随机数，Monte Carlo 方法解决几类问题.尽管在理论上这些方法最终将收敛于正确解，但在实际上，仅达到适当的精度.这是因为收敛速度极慢.有时 Monte Carlo 方法用来对加速改进算法求得好的起始值.提供两类应用.

1. 模拟是指对"实际"现象提供算法模拟的方法.在广义上这描述了应用数学的一般概念.例如微分方程可以模拟导弹的飞行.但是在这里模拟这个术语是指 Monte Carlo 方法中随机过程中的模拟.经典的例子是中子向反应器的屏障运动的模拟.它的曲折的途径为算术随机游动所模拟.(见题 30.2 和 30.4)
2. 抽样是指通过研究小的随机子集，推出一个大集合元素的性质的方法.于是 $f(x)$ 在一个区间上的平均值能通过在这区间内点的有限随机子集的平均值来估计.因为 $f(x)$ 的平均实际上是一个积分.这就相当于 Monte Carlo 方法和近似积分.作为第二个例子，单位圆上一组 N 个随机点的重力中心的测定可以用几百个或几千个这种集合作样本来研究.(见题 30.5)

题 解

30.1 什么是随机数，它们是如何生成的?

解 ☞ 作为一个简单却有益的例子是从数 01 开始,乘以 13 得到 13,再乘以 13 并去掉百位数得到 69,以这种方法进行下去,继续乘以 13 除去模 100 产生了以下两位数字的序列

01,13,69,97,61,93,09,17,21,73,49,37,81,53,89,57,41,33,29,77.

77 以后,序列再从 01 开始.

已产生这些数字的方法毫不随机,但是这些数字还是被称为随机数.假如我们把它们放到从 00 到 99 的刻度尺上,它们显得是相当均匀地分布,对刻度尺的任一部分没有一点偏爱.从 01 开始再返回,连续地取这些值,我们发现 10 次增加 10 次减少,将它们取成三个一组,我们发现两次增加(诸如 01,13,69)与两次减少一起出现大约一半的时间.正符合概率论理论所提示的.随机数这个术语是应用在一串数上,它通过合理次数的随机性的这类概率试验.当然我们的序列太短对任何有经验的试验还站不住脚.假如我们数出三个增加(诸如 01,13,69,97)与三个减少一起,我们发现比它们本该如此的还多.所以我们不能期望太高.正如它的本来面目,这个序列比我们以 5 为乘数所得到的还要好一些,(01,05,25,25,25.…它们在任何意义下都不是随机数),一个小的乘数诸如 3 会导致(01,03,09,27,81,…)这一直向增加前进,简直不是一个好兆头,看看好好地选择一个大的乘数,可能是最好的.

30.2 在中子穿过原子反应堆的铅屏障运动的模拟中应用前问题中的随机数.

解 ☞ 为了简单,我们假设每一个进入屏障的中子在撞到铅原子前,行进距离 D,然后这个中子以随机方向弹回来,并且在它的下一次撞击中又行进距离 D.再假设屏障的厚度是 $3D$,尽管这个厚度不足以适当的防护.最后假设所有的中子每一个能经受十次撞击.进入的中子有多少比例能穿透这铅屏障?假如我们的随机数被解释为方向(图 30.1),那么它们可以用来预测弹回的随机方向.假如从 01 开始将得到图30.2中虚线表示的路线.这个中子在四次撞击后就穿过了.第二个中心沿着图 30.2中的实线在十次撞击后停在屏障内.现在明白对实际企图我们没有足够随机数.不过可参见题 30.3.

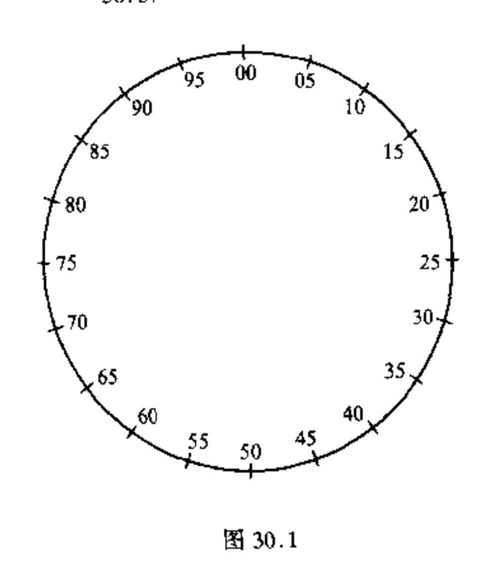


图 30.1

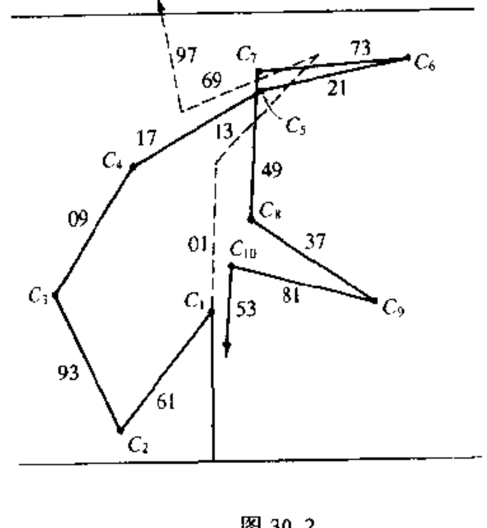


图 30.2

30.3 如何能产生更广泛的随机数供应?

解 ☞ 现在有相当多的方法,但是其中最好的是用题 30.1 中的模乘想法.例如,递推公式

$$x_{n+1} = 7^9 x_n (\mathrm{mol}\ 10^s), \quad x_0 = 1$$

产生了有很好的统计性质,长度为 $5 \cdot 10^{s-3}$的序列.它适合于十进制机器.递推公式

$$x_{n+1} = (8t-3)x_n (\mathrm{mod}\ 2^s), \quad x_0 = 1$$

产生了有一定统计性质的一列数 $1,5,9,\cdots,2^s-3$,它适合二进制机器.数 t 是任意的,但应选大的以免向上的过程太长.在这两种方法中,s 代表计算机的标准字长,在十进制机器中多半 $s=10$.二进制机器多半 $s=3\pi$.

30.4 用一种好的随机数供应继续题 30.2.

解 ☞ 在十位数字机器($s=10$)上用题 30.3 中第一个序列,就得到下面给出的结果.这些是 Monte Carlo 方法的典型结果,朝正确解答的收敛很慢.好象有 28% 将穿透铅墙,因此必须用更厚的铅屏障.

试验数	5,000	10,000	15,000	20,000
穿透百分比	28.6	28.2	28.3	28.4

30.5 假设在单位圆周上随机地选取 N 个点,我们能预期其重力中心降落何处?

解 ☞ 根据对称性,其重力中心的角坐标应均匀分布,即角的位置彼此相同,它的径坐标更有趣,我们通过抽样技术来逼近它.题 30.3 中的序列中,每个随机数前面可加上十进制小数点(或二进制),并且乘以 2π.结果是 0 到 2π 之间的随机角 θ_i,我们用它来测定单位圆上的随机点.把 N 个这种随机点放在一起,它们的重力中心在

$$X=\frac{1}{N}\sum_{i=1}^{N}\cos\theta_i,\quad Y=\frac{1}{N}\sum_{i=1}^{N}\sin\theta_i,$$

径坐标是 $r=\sqrt{X^2+Y^2}$.把区间 $0\leqslant r\leqslant 1$ 分成长为 $\frac{1}{32}$ 的子区间.下面我们提示这个特殊的 r 值落在哪个子区间.于是取 N 个随机点新的取样,并重复这个过程.用这种方法,我们得到了径坐标分布的随机近似.6000 多种取样,对 $N=2,3$ 和 4 的情形的结果在表 30.1 中给出.冠以 Freq 的列给出了重力中心出现在每个子区间(从中心向上)的真正的频率.冠以 Cum 的列给出了累积的部分,对于 $N=2$ 的情形,这累积结果还碰巧刚好是 $(2/\pi)\arcsin(r/2)$.它作为一种精度检查,注意我们大约有三位精度.

表 30.1

	$n=2$			$n=3$		$n=4$	
	Freq	Cum	Exact	Freq	Cum	Freq	Cum
1	121	0.0197	0.0199	7	0.001	36	0.005
2	133	0.0413	0.0398	37	0.007	87	0.018
3	126	0.0618	0.0598	58	0.017	128	0.038
4	124	0.0820	0.0798	67	0.028	169	0.063
5	129	0.1030	0.0999	95	0.043	209	0.094
6	111	0.1211	0.1201	113	0.061	192	0.123
7	123	0.1411	0.1404	141	0.084	266	0.163
8	115	0.1598	0.1609	172	0.112	289	0.207
9	129	0.1808	0.1816	224	0.149	238	0.242
10	142	0.2039	0.2023	336	0.203	316	0.290
11	123	0.2240	0.2234	466	0.279	335	0.340
12	138	0.2464	0.2447	344	0.335	360	0.394
13	126	0.2669	0.2663	291	0.383	357	0.448
14	157	0.2925	0.2883	285	0.429	365	0.503
15	126	0.3130	0.3106	269	0.473	365	0.558
16	125	0.3333	0.3333	255	0.514	405	0.618
17	150	0.3577	0.3565	223	0.551	353	0.672

续表

	$n=2$			$n=3$		n 4	
	Freq	Cum	Exact	Freq	Cum	Freq	Cum
18	158	0.3835	0.3803	189	0 581	255	0.710
19	135	0.4054	0.4047	208	0.615	275	0.751
20	148	0.4295	0.4298	185	0.645	262	0.790
21	157	0.4551	0.4558	215	0.680	182	0.818
22	158	0.4808	0.4826	197	0.712	159	0.842
23	173	0.5090	0.5106	183	0.742	163	0.866
24	190	0.5399	0.5399	201	0.775	168	0.892
25	191	0 5710	0.5708	188	0.805	167	0.917
26	211	0.6053	0.6038	183	0.835	131	0.936
27	197	0.6374	0.6393	163	0.862	102	0 952
28	247	0.6776	0.6783	176	0.890	87	0.965
29	262	0.7202	0.7221	170	0.918	87	0.978
30	308	0.7703	0.7737	162	0.944	76	0.989
31	424	0.8394	0.8407	163	0.971	45	0.996
32	987	1.0000	1.0000	178	1.000	27	1.000

30.6 通过使用随机游动的取样方法求解边值问题

$$T_{xx}+T_{yy}=0,\quad T(0,y)=T(1,y)=T(x,1)=0,\quad T(x,0)=1.$$

解 这是一个不带有明显统计风味的问题.它可以转化为适合于 Monte Carlo 方法的形式.熟悉的有限差分近似导出一组离散的点(在图 30.3 中九点),在每一个这样的点上下面的方程

$$T_5=\frac{1}{4}(T_2+T_4+T_6+T_8)$$

使得每一个 T 值是它的四个邻点上 T 值的平均.这种九个方程的相同集合在题 26.29 中曾遇到过,每个未知量代表迷路的狗最终出现在我们图中南面的边上的概率.也可重新解释为回廊迷宫.尽管取样近似在这里不是最经济的,但是了解它是如何进行的是很有趣的.例如一条虚构的狗在位置 1 出发,我们产生了一个随机数.根据这一随机数在子区间 $\left(0,\frac{1}{4}\right)$,$\left(\frac{1}{4},\frac{1}{2}\right)$,$\left(\frac{1}{2},\frac{3}{4}\right)$,或者 $\left(\frac{3}{4},1\right)$ 中的哪一个,我们的狗从北方,东方,南方或者西方朝下一个交点运动,我们检查看看这样是否把它带出迷宫.如果没有,就产生了另一个随机数,并且进行第二次运动.当狗最后出现在某处,我记录这是否是南面的边或者不是.然后,在位置 1 处的新的虚构狗又出发,并重复这种行动 10 000 次这种计算机取样的结果有 695 次成功地在南面出口处出现.这就使得成功的概率是 0.0695.这应该和用 Gauss-Seidel 迭代求得的结果 0.071 进行比较.后者是更精确.但是用取样方法解微分方程两点边值问题的可能性在更复杂情形下可能是有用的.

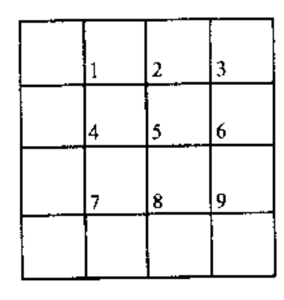


图 30.3

30.7 叙述用 Monte Carlo 方法进行积分的近似计算.

解 最简单的方法可能是通过平均来求积分的近似值.

$$\int_a^b f(x)\mathrm{d}x\approx\frac{1}{N}\sum_{i=1}^{N}f(x_i)$$

这里 x_i 是(a,b)中随机地选取.例如,假如我们刚好使用题 30.1 中的最先五个随机数,每一个在前面加上小数点,于是我们有

$$\int_0^1 x\mathrm{d}x\approx\frac{1}{5}(2.41)\approx 0.48,$$

而这个正确值是$\frac{1}{2}$.我们还求得 $\int_0^1 x^2dx\approx 0.36$,而它的正确值是$\frac{1}{3}$.对于同样的积分,当 $N=100$,

用题30.3中较长的序列,结果分别得到0.523和0.316,误差大约是5%.这不太精确,但在高维积分的情形有同样的精度,而且 Monte Carlo 方法比其他积分算法好.

补 充 题

30.8 用 $x_{n+1}=rx_n \pmod{100}$生成20个随机数的序列,你自己选取乘子 r.像在题30.2中那样,用这些数去模拟三条或四条中子的路线.

30.9 用题30.3的那种序列,像在题30.4中那样1000个中子路线.对厚度为$5D$,$10D$和$20D$的铅屏障重复进行、防护效应是怎样增加的?

30.10 模拟在平面上的1000个点的随机运动,每一次运动是25步,每步的长度相等.设每次运动在(0,0)点开始,而且每步是随机方向.计算从(0,0)点开始到4,9,6和25步后的平均距离.

30.11 用随机数近似计算$\int_0^\pi \sin x\,dx$

30.12 用随机数近似计算

$$\int_0^1\int_0^1\int_0^1\int_0^1\int_0^1\int_0^1 \frac{dA\,dB\,dC\,dD\,dE\,dF}{1+A+B+C+D+E+F}.$$

30.13 高尔夫球手A和B有以下记录

得分	80	81	82	83	84	85	86	87	88	89
A	5	5	60	20	10					
B				5	5	10	40	20	10	10

A行和B行的数字表示每人击中所给线号的次数.假设他继续这种性质的游戏,而且A允许B每轮四次击中(意思是B可从他的记分中减去四次击中).模拟这两人间的1000次比赛.A多久打败B一次?他们多久打平手一次?

30.14 A,B和C每人有普通的一包卡片,他们进行了洗牌并且每人随机地摊开一张卡片.三张卡片显示的可以包含1,2或3不同的花式,赢者决定如下:

显示花式的数字	1	2	3
赢者是	A	B	C

取代摊开了的卡片就完成了这游戏,假如进行很多次这种游戏,每人多久赢一次?初等概率能求出每个答案,但是通过每次产生三个随机数来模拟真实的游戏,按下面的格式决定花式:

x落到区间内	$\left(0,\frac{1}{4}\right)$	$\left(\frac{1}{4},\frac{1}{2}\right)$	$\left(\frac{1}{2},\frac{3}{4}\right)$	$\left(\frac{3}{4},1\right)$
花式是	S	H	D	C

30.15 棒球手在一次比赛中平均0.300轮到四次做打手,他分别击中0,1,2,3和4次的机会是多少?能用初等概率求得答案,但是要用模拟求解.

30.16 在"回到零的第一人"游戏中,两个参与者轮流把同一个指标点向后或向前移动通过棋盘.

10	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10

指标点在0开始,参与者A开始并且一直向右运动而B向左运动.运动的正方形的数字是由投骰子决定的.停在0处的第一个人就是赢者,假如指标点跑出棋盘的任一端,游戏为平局,指标点回到0,

同时由参与者 A 开始新的游戏. A 赢的可能性有多大? 用概率论不太容易求得答案, 用模拟方法进行求解.

30.17　把整数 1 到 N 进行随机地排列. 没有整数在它的自然位置的可能性有多大? 这就是著名的"重数问题", 而且用概率论解决了. 不过在这里可选取某些 N 的值用模拟方法求解.

30.18　产生三个随机数. 按照增加的次序 $x_1<x_2<x_3$ 排列它们, 重复许多次, 并且计算 x_1 的平均值, x_2 的平均值和 x_3 的平均值.

30.19　假设要求随机数 y 非均匀分布, 密度是 $f(y)$. 这种数 y 能由均匀分布的随机数 x 通过使两者的累积分布相等而产生, 即

$$\int_0^x 1\cdot \mathrm{d}x = \int_0^y f(y)\mathrm{d}y.$$

对于特殊情形 $f(y)=e^{-y}$, 说明从 x 可计算出 y 是多少?

30.20　对于正态分布 $f(y)=e^{-y^2/2}/\sqrt{2\pi}$, 前一问题中的方法比较麻烦. 通常可供选择的方法是从(0, 1)的均匀分布产生 12 个随机数 x, 再对它们求和并减去 6, 这是因为常常用零的平均值代替正态分布较佳. 这一过程是依赖以下事实, 即若干个均匀分布的随机数的和是接近于正态分布的. 用它来产生 100 或 1000 个随机数

$$y = \left(\sum_{i=1}^{12} x_i\right) - 6,$$

然后再检查产生的数 y 的分布. 它们中间多少是在区间(0, 1), (1, 2), (2, 3)和(3, 4)内? 相应的负区间应有类似的份额.

补充题答案

第一章

1.39　$1+0.018$，仅需两项.

1.40　-0.009.

1.41　$N=100, N=10,000$.

1.42　0.114904, 0.019565, 0.002486, 0.000323, 0.000744, 0.008605.

1.43　0.008605.

1.44　算出 $J_8=0.119726$.

1.48　0.1494 近似.

1.49　超过$\frac{15}{8}$，上溢；$\frac{1}{4}$以下，下溢.

1.56　二进制的点，近似.

1.57　L_1 适合于出租车，L_∞适合于国王.

第二章

2.11　$(x-1)(x^2+1)$.

2.12　$3, -3, 3, -3, 3$.

2.13　$p(x)=2x-x^2$.

2.15　最大的估计误差 $=0.242$；实际误差 $=0.043$.

2.16　$y'=1.11, p'=1$.

2.17　$y''=-1.75, p''=-2$.

2.18　$4/\pi, \frac{4}{3}$.

2.19　$y=x+7x(x-1)+6x(x-1)(x-2)$.

2.20　$\pi(x)=x(x-1)(x-2)(x-3)$.

2.21　1.

第三章

3.13　4 阶差分全为 24.

3.14　$\Delta^5 y_0=\Delta^4 y_1-\Delta^4 y_0$ 并且，现用我们对于 4 阶差分的结果.

3.15　$\frac{u_{k+1}}{v_{k+1}}-\frac{u_k}{v_k}=\frac{v_k u_{k+1}-u_k v_{k+1}}{v_{k+1}v_k}$，等.

3.16　5 阶差分是 $5, 0, -5$.

3.17　将 y_2 改为 0.

3.22　1, 3, 7, 14, 25, 41.

3.23　$\Delta y_k=0, 1, 5, 18, 36, 60$；$y_k=0, 0, 1, 6, 24, 60, 120$.

3.24　$\Delta^2 y_k=24, 30, 36$；$\Delta y_k=60, 90, 126$；$y_k=120, 210, 336$.

3.25　将 113 改为 131.

3.26　$\Delta^2 y_1=y_3-2y_2+y_1$；$\Delta^2 y_2=y_4-2y_3+y_2$.

3.27　3^k.

3.28　$4^k, (-2)^k$.

3.29　$\frac{1}{6}[4^k \quad (-2)^k]$.

3.30　对差分的正弦使用恒等式.

3.31　对差分的余弦使用恒等式.

第四章

4.23 $120, 720, 0, -\frac{2}{9}, \frac{10}{27}, -\frac{80}{81}$.

4.24 $\frac{1}{7}, \frac{1}{56}, \frac{1}{504}, \frac{3}{4}, \frac{9}{28}, \frac{27}{280}$.

4.25 $20, 1, 0, -\frac{1}{9}, \frac{5}{81}, -\frac{10}{243}$.

4.26 4 阶差分全为 24.

4.27 $4k^{(3)}, 12k^{(2)}, 24k, 24$.

4.28 $5k^{(4)}, 20k^{(3)}, 60k^{(2)}, 120k, 120$.

4.29 $2k^3 - 7k^2 + 9k - 7$.

4.30 $k^6 - 15k^2 + 85k^4 - 224k^3 + 271k^2 - 118k + 1$.

4.31 $\frac{2}{3}k^{(4)} + 4k^{(3)} + 2k^{(2)} - 2k^{(1)} + 1$.

4.32 $3k^{(5)} - 25k^{(3)} + 75k^{(2)} + 53k^{(1)}$.

4.33 $\Delta y_k - 53 + 135k + 90k^2 - 90k^3 + 15k^4$.

4.34 $\Delta^2 y_k = 150 - 30k - 180k^2 + 60k^3$.

4.35 31, 129, 351.

4.36 10, 45, 126.

4.37 2.

4.38 4.

4.39 $k^{(3)}/3$.

4.40 $k^{(4)}/4$.

4.41 $\frac{1}{3}k^{(3)} + \frac{1}{2}k^{(2)}$.

4.42 $\frac{1}{2}k^{(2)} + k^{(3)} + \frac{1}{4}k^{(4)}$.

4.43 $-1/(k+1)$.

第五章

5.9 $\frac{1}{2}[n + 1^{(2)} - 1^{(2)}]$.

5.10 $n^2(n+1)^2/4$.

5.11 利用该事实 $A^i = \Delta[A^i/(A-1)]$.

5.12 利用该事实 $\binom{i}{k} = i^{(k)}/k! \ = \Delta[i^{(k+1)}/(k+1)!]$.

5.13 $\frac{1}{4}$.

5.14 $\frac{3}{4}$.

5.15 $(R^3 + 4R^2 + R)/(1-R)^4$.

5.16 26.

5.17 $-\frac{1}{3}$.

5.18 $\log(n+1)$.

5.19 $\sum_{j=1}^{n} |s_j^{(n)}[(N+1)^{(j+1)}]/(j+1)|$.

5.20 $\frac{1}{n}\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right)$.

5.21 用 $S_n(R)$表示该和，则 $S_{n+1}(R) = RS'_n(R)$可用来依次计算每一个和.

5.22 $y_k = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k-1}$.

5.23 $y_k = \log 2 + \log 3 + \cdots + \log(k-1)$.

第六章

6.8　$[(x-2)(x-4)/64][8-4(x-6)+(x-6)(x-8)]$.

6.9　$1+x+\frac{1}{2}x(x-1)$.

6.10　$6+18(x-3)+9(x-3)(x-4)+(x-3)(x-4)(x-5)$.

6.11　4 次足够了，$x(x-1)\left[\frac{1}{2}-\frac{1}{3}(x-2)+\frac{1}{12}(x-2)(x-3)\right]$.

6.12　$1+x+\frac{1}{2}x(x-1)+\frac{1}{6}x(x-1)(x-2)$.

6.14　$7x^2-6x$.

6.15　$\frac{1}{3}x^3-2x^2+\frac{8}{3}x$；在 $x=4$ 上配置，但在 $x=5$ 上不配置.

6.16　否，3 次.

6.17　否，一次.

6.18　$(7x^2-x^4)/6$；在 $(-2,-1)$ 与 $(1,2)$ 内较大.

6.19　$(7x-x^2)/6$；自变量不等距.

6.20　$y_k=\frac{1}{6}k(k-1)/(k-2)$.

第七章

7.33　$1+2k+2k(k+1)+\frac{4}{3}k(k+1)(k+2)+\frac{2}{3}k(k+1)(k+2)(k+3)$.

7.34　$120+60k+12k(k+1)+k(k-1)(k+2)$.

7.36　$2x-3x^2+x^3$.

7.37　$1-k-k(k-1)+\frac{1}{2}(k+1)k(k-1)+\frac{1}{4}(k+1)k(k-1)(k-2)$.

7.38　$1+k-(k+1)k-\frac{1}{2}(k+1)k(k-1)+\frac{1}{4}(k+2)(k+1)k(k-1)$.

7.39　$24+36k+9k(k-1)+(k+1)k(k-1)$.

7.40　$1-\frac{1}{2}k(k-1)+\frac{1}{12}(k+1)k(k-1)(k-2)$.

7.42　$1-k^2+\frac{1}{4}(k+1)k^2(k-1)$.

7.43　当 $x=1$ 时 $k=0$，$y=2+\frac{3}{2}k+\frac{1}{2}k^2$.

7.44　$60k-24(k-1)+4(k+1)k(k-1)-3k(k-1)(k-2)$.

7.45　$1-\frac{1}{6}[(k+1)k(k-1)-k(k-1)(k-2)]+\frac{1}{60}[(k^2-4)(k^2-1)k-(k^2-1)k(k-2)(k-3)]$.

7.46　$4k-2(k-1)+\frac{1}{6}[(k^2-1)k-k(k-1)(k-2)]$.

7.47　$42+36\left(k-\frac{1}{2}\right)+\frac{21}{2}k(k-1)+\left(k-\frac{1}{2}\right)k(k-1)$.

7.48　$1-\frac{1}{2}k(k-1)+\frac{1}{12}(k-1)k(k-1)(k-2)$.

第八章

8.15　$\frac{(x-1)(x-4)(x-6)}{-24}-\frac{x(x-4)(x-6)}{15}+\frac{x(x-1)(x-6)}{-24}$
$-\frac{x(x-1)(x-4)}{60}$；$y(2)=-1$，$y(3)=0$，$y(5)=1$.

8.16　$-\frac{4x(x-2)(x-4)(x-5)}{3}+4x(x-1)(x-4)(x-5)$
$-11\frac{x(x-1)(x-2)(x-5)}{3}$；$y(3)=84$.

8.18　$a_0=\frac{5}{2}$；$a_1=\quad 15$；$a_2=\frac{31}{2}$.

8.19　$\frac{\frac{2}{35}}{x+1}+\frac{\frac{4}{15}}{x-1}-\frac{\frac{41}{30}}{x-4}+\frac{\frac{73}{70}}{x-6}$.

8.22 $\frac{(x-x_1)^2}{(x_0-x_1)^2}\left[\left(1-\frac{2(x-x_0)}{x_0-x_1}\right)y_0+(x-x_0)y_0\right]$
$+\frac{(x-x_0)^2}{(x_1-x_0)^2}\left[\left(1-\frac{2(x-x_1)}{x_1-x_0}\right)y_1+(x-x_1)y_1'\right].$

8.23 1 阶, $-2,\frac{2}{3},-1$;2 阶, $\frac{2}{3},-\frac{1}{3}$;3 阶, $-\frac{1}{6}$.

8.24 $1-2x+\frac{2}{3}x(x-1)-\frac{1}{6}x(x-1)(x-4)$.

8.25 1 阶, $\frac{2}{3},0,-\frac{1}{3}$;2 阶, $-\frac{1}{3},\frac{1}{3}$;3 阶, $-\frac{1}{6}$.

8.26 -1.

8.27 $16x+8x(x-1)-3x(x-1)(x-2)-x(x-1)(x-2)(x-4);y(3)=84$.

第九章

9.22 $C_0=C_4=0,C_1=C_3=\frac{18}{7},C_2=-\frac{30}{7}$.

9.23 $S_2(x)=(2-x)^3/6-7(x-1)^3/12-(2-x)/6+19(x-1)12$;
$S_3(x)=-7(3-x)^3/12+(x-2)^3/6+19(3-x)/12-(x-2)/6$.

9.24 d_i 全为零.

9.25 d_i 为 y 的二阶均差的 6 倍,除端点条件外所有方程还原为 $3c=d_i$.

第十章

10.8 $2x^2-x^3$.

10.9 $x^4-4x^3+4x^2$.

10.10 $3x^5-8x^4+6x^3$.

10.11 $p_1(x)=\frac{1}{4}x^2,p_2(x)=2-\frac{1}{4}(4-x)^2$.

10.12 $p_1(x)=x^3(4-x)/16,p_2(x)=2-(4-x)^3x/16$.

10.15 x^4-2x^2+1.

10.16 $2x^4-x+1$.

10.17 x^3-x^2+1.

第十一章

11.20 $\sin x=x-x^3/3!+x^5/5!-x^7/7!+\cdots$对于奇次 n;
$\cos x=1-x^2/2!+x^4/4!-x^6/6!+\cdots$对于偶次 n.

11.21 $\pm\sin\xi\cdot x^{n+1}/(n+1)!$ 对于这 2 个函数.

11.22 $n=7$.

11.23 $n=8,n=12$.

11.24 $\sum_{i=1}^{\infty}D^i/i!$.

11.27 $\delta+\frac{1}{2}\delta^2+\frac{1}{8}\delta^3-\frac{1}{128}\delta^5+\frac{1}{1024}\delta^7+\cdots$.

第十二章

12.31 1.0060, 1.0085, no.

12.32 1.0291.

12.33 1.01489.

12.34 1.12250.

12.35 1.05830.

12.36 0.12451559.

12.37 0.1295.

12.38 1.4975.

12.39　1.4975.

12.40　0.1714, 0.1295, 0.0941.

12.41　0.02.

12.42　0.006.

12.43　0.25, 0.12.

12.45　大约 1.

12.48　对于 $x>1$ 大约 $h=0.15$.

12.49　$\frac{5}{4}$.

12.51　15.150.

12.52　0.841552021.

12.54　1.16190, 1.18327, 1.20419, 最后一项舍去 3 个单位.

12.55　1.20419, 1.22390, 两者多少带有误差.

12.56　误差 $=x^4-7x^2+6x$; $\xi=0$ 解释误差为 0.

12.57　幸运的 ξ 值.

12.58　0.

12.59　24.

12.60　0 与 1.

第十三章

13.22　$hp'=\delta y_{1/2}+\left(k-\frac{1}{2}\right)\mu\delta^2 y_{1/2}+\frac{6k^2-6k+1}{12}\delta^3 y_{1/2}+\frac{4k^3-6k^2-2k+2}{24}\mu\delta^4 y_{1/2}$

$+\frac{5k^4-10k^3+5k-1}{120}\delta^5 y_{1/2}$,

$h^2 p^{(2)}=\mu\delta^2 y_{1/2}-\left(k-\frac{1}{2}\right)\delta^3 y_{1/2}+\frac{12k^2-12k-2}{24}\mu\delta^4 y_{1/2}+\frac{4k^3-6k^2+1}{24}\delta^5 y_{1/2}$,

$h^3(p)^{(3)}=\delta^3 y_{1/2}+\left(k-\frac{1}{2}\right)\mu\delta^4 y_{1/2}+\frac{1}{2}(k^2-k)\delta^5 y_{1/2}$,

$h^4 p^{(4)}=\mu\delta^4 y_{1/2}+\left(k-\frac{1}{2}\right)\delta^5 y_{1/2}$　$h^5 p^{(5)}=\delta^5 y_{1/2}$.

13.23　0.4714, −0.208, 0.32.

13.24　预测近似误差 10^{-9}; 实际误差 0.000038.

13.25　最大舍入误差约为 $2.5E/h$; 对于表 13.1 它变成 0.00025.

13.28　精确结果为 $x=\pi/2$, $y=1$.

13.29　1.57

13.31　$h^5=3E/8A$; $h\approx 0.11$.

第十四章

14.41　$h\approx\sqrt{3}/100$.

14.42　$A_2=0.69564$, $A_1=0.69377$, $(4A_1-A_2)/3=0.69315$.

14.43　0.69315.

14.44　0.6931, 无须校正.

14.45　$h=0.14$.

14.46　梯形公式为 $\sqrt{3}/10^4$; Simpson 公式为 0.014.

14.52　精确值为 $\pi/4=0.7853982$.

14.53　准确值为 1.4675.

14.58　0.36422193.

14.60　9.68848.

14.62　$a_{-1}=a_1=\frac{7}{15}$, $a_0=\frac{16}{15}$, $b_0=0$, $b_{-1}=-b_1=\frac{1}{15}$.

14.67　0.807511.

第十五章

15.56 1.0000081.

15.57 1.5.

15.61 $L_0=1, L_1=1-x, L_2=2-4x+x^2, L_3=6-18x+9x^2-x^3$,
$L_4=24-96x+72x^2-16x^3+x^4, L_5=120-600x+600x^2-200x^3+25x^4-x^5$.

15.68 精确值是 5.

15.69 准确到 5 位是 0.59634.

15.71 $H_0=1, H_1=2x, H_2=4x^2-2, H_3=8x^3-12x, H_4=16x^4-48x^2+12, H_5=32x^5-160x^3+120x$.

15.73 $(\sqrt{\pi}/6)\left[y\left(-\sqrt{\frac{3}{2}}\right)+y\left(\sqrt{\frac{3}{2}}\right)+4y(0)\right]; 3\sqrt{\pi}/4$.

15.77 2.128.

17.78 0.587.

15.80 2.404.

15.81 3.82.

第十六章

16.13 0.5 与 −0.23，与精确值 0.5 及 −0.25 相比.

16.15 1.935.

16.18 −0.797.

第十七章

17.50 $n(n-1)(n-2)/3$.

17.51 $(n+1)^2n^2(2n^2+2n-1)/12$.

17.52 $\frac{3}{4}-\frac{2n+3}{2(n+1)(n+2)}$.

17.55 $\frac{11}{18}-\frac{1}{3}\left(\frac{1}{n+1}+\frac{1}{n+2}+\frac{1}{n+3}\right)$.

17.57 0.6049.

17.61 大约 $x=0.7$.

17.62 至多为 8.

17.63 大约 $x=0.7$.

17.64 $\frac{x^{2n+1}}{(2n+1)!}\cdot\frac{(2n+2)^2}{(2n+2)^2-x^2}$；大约 $x=10$.

17.65 0.798.

17.66 0.687.

17.67 0.577.

17.68 1.1285.

17.73 $Q_t=x^t$.

17.78 在 4 项之后；此法提供 $C\approx0.5769$.

17.86 在 7 项之后.

第十八章

18.31 $y_k=\left[A+\frac{1}{(1-r)^2}\right]r^k+\frac{k}{1-r}-\frac{1}{(1-r)^2}$, $r=1$ 除外.

18.32 1, 3, 1, 3, etc.; $2-(-1)^k$; $(y_0-2)(-1)^k+2$.

18.35 令 $y_k=(k-1)!\ A(k)$ 当 $k>0$，得到 $y_k=(k-1)!\ (2^k-1)$.

18.36 $\frac{127}{64}$.

18.37 $\left(\left(\left(\left(\frac{x^2}{9\cdot8}-1\right)\frac{x^2}{7\cdot6}+1\right)\frac{x^2}{5\cdot4}-1\right)\frac{x^2}{3\cdot2}+1\right)x$.

18.40　$1/(k-1)!$.

18.41　$\psi^{(3)}(0)=3!\pi^4/90$, $\psi^{(3)}(n)=3!\left[\frac{\pi^4}{90}-\sum_{k=1}^{n}\frac{1}{k^4}\right]$.

18.42　$\frac{3}{4}$.

18.43　$\pi^2/12-\frac{11}{16}$.

18.44　$\psi\left(\frac{1}{2}\right)=0.0365$, $\psi\left(\frac{3}{2}\right)=0.7032$, $\psi\left(-\frac{1}{2}\right)=1.9635$.

18.45　取任意大的负值.

18.46　$\frac{2}{3}\psi(0)-\frac{1}{3}\psi\left(\sqrt{\frac{3}{5}}\right)-\frac{1}{3}\psi\left(-\sqrt{\frac{3}{5}}\right)$.

18.47　$\frac{1}{3}\psi(0)-\frac{1}{6}\psi\left(\sqrt{\frac{3}{4}}\right)-\frac{1}{6}\psi\left(-\sqrt{\frac{3}{4}}\right)$.

18.50　$5(-1)^k-3(-2)^k$.

18.52　$A+B(-1)^k$.

18.53　$A4^k+B3^k+(a\cos k+b\sin k)/(a^2+b^2)$,

其中 $a=\cos 2-7\cos 1+12$, $b=\sin 2-7\sin 1$,

$A=(3a-a\cos 1-b\sin 1)/(a^2+b^2)$,

$B=(-4a+a\cos 1+b\sin 1)/(a^2+b^2)$.

18.54　$\left[-4\left(-\frac{1}{2}\right)^k+2k\left(-\frac{1}{2}\right)^k+3k^2-8k+4\right]\Big/27$.

18.56　$\frac{2}{3}\left[2^k-\left(\frac{1}{2}\right)^k\right]$.

18.57　$\left[5^k\left(-\cos k\theta-\frac{5}{4}\sin k\theta\right)+2^k\right]/41$, $\cos\theta=-\frac{3}{5}$, $\sin\theta=\frac{4}{5}$.

18.59　$a<0$.

18.60　$\frac{1}{8}(3^k)-\frac{1}{16}(-1)^k-\frac{3}{8}k^2-\frac{1}{16}$.

18.61　振荡的,线性的,指数的.

18.65　$\frac{1}{2}[1-(-1)^k]$.

第十九章

19.76　精确值是 1.

19.77　1.4060059.

19.78　精确解为 $x^3y^4+2y=3x$.

19.79　精确解为 $x^2y+xe^y=1$.

19.80　精确解为 $\log(x^2+y^2)=\arctan y/x$.

19.81　4 天,18 小时,10 分.

19.82　4.

19.83　精确值为 $\frac{1}{8}\arctan\frac{1}{4}$.

19.84　精确解为 $x=-\sqrt{1-y^2}+\log\left(1+\sqrt{1-y^2}\right)/y$.

第二十章

20.16　参看题 19.87.

20.19　$a_0=a_1=1$, $k^2a_k-(2k-1)a_{k-1}+a_{k-2}=0$ 当 $k>1$.

20.20　对 e^{-21h}的 4 次 Taylor 近似为 6.2374 与准确值 0.014996 相比较.

第二十一章

21.57　$y=0.07h+4.07$.

21.58　4.49,4.63,4.77,4.91,5.05,5.19,5.33,5.47,5.61,5.75.

21.59 0.07.

21.60 否.

21.62 很小.

21.63 它们是交替的.

21.65 $A = 84.8, M = -0.456$.

21.67 这里5点公式确实较好.

21.69 其结果几乎与五点公式的相同.

21.85 $p(x) = \frac{1}{3}$.

21.86 $p(x) = 3x/5$.

21.87 $p(x) = 3x/5$.

21.88 $p(x) = 0.37 + 0.01x - 0.225(3x^2 - 1)/2$.

21.90 $p(x) = \frac{1}{2}$.

21.91 $p(x) = 3x/4$.

21.92 去掉两项我们得到 $1.2660T_0 - 1.1303T_1 + 0.2715T_2 - 0.0444T_3 + 0.0055T_4 - 0.0005T_5$.

21.102 $(81 + 75x)/64$;在$(-1,1)$上它仅略差于二次的.

21.106 $3x/4$.

21.107 极小,积分抛物线为 $p = \frac{2}{\pi} + \frac{4}{3\pi}(2x^2 - 1)$.

21.109 0.001,0.125,0.217,0.288,0.346,0.385,0.416,0.438,0.451,0.459,0.466.

21.110 −0.8,19.4,74.4,143.9,196.6,203.9,108.2,143.4,126.7,118.4,112.3,97.3,87.0,73.3,56.5,41.8,33.4,26.5,15.3,6.6,1.2.

21.111 $5.045 - 4.043x + 1.009x^2$.

第二十二章

22.34 $P = 4.44e^{0.45x}$.

22.37 $p = \frac{5 - 3\sqrt{3}}{16} + \frac{3}{\pi}\left(\sqrt{3} - \frac{1}{2}\right)x + \frac{9}{\pi^2}\cdot\frac{1-\sqrt{3}}{2}x^2$.

22.38 $p = (1 - 18x + 48x^2)/32; h = \frac{1}{32}$.

22.41 $(10T_0 + 15T_2 + 6T_4)/32; \frac{1}{32}$.

22.42 $T_0 + T_1 + T_2; 1$.

22.43 $\frac{1763}{2304}T_0 - \frac{353}{1536}T_2 + \frac{19}{3840}T_4; 1/23,040$.

22.44 $p = 2x/\pi - 1.10525$.

22.45 方法失败,x_2 成为间断点.

22.46 $p = -2x/\pi + 1.105$.

22.50 $1.6476 + 0.4252x + 0.0529x^2; 0.0087$.

22.51 4次.

22.52 不超过0.000005.

22.53 4次.

22.54 2次.

第二十三章

23.18 $3/x$;否,该方法得出 $4 - x$.

23.19 $90/(90 + 97x - 7x^2)$;否,该方法得出$(20 + 7x)/(20 + 34x)$.

23.20 $(x^2 - 1)/(x^2 + 1)$.

23.21 $x^2/(1 + x)$.

23.22 $(x + 1)/(x + 2)$.

23.24 $1/(2 - x^2)$.

23.25　$-\frac{1}{2}$.

23.28　$4(1-x+x^2)/(1+x)$.

23.29　$12(x+1)/(4-x^2)$.

23.30　$(x^2+x+2)/(x^2+x+1)$.

23.31　$1/(\sin 1°30') \approx 38.201547$.

23.32　$(1680-2478x+897x^2-99x^3)/(140+24x-17x^2)$.

23.33　$(24+18x+6x^2+x^3)/(24-6x)$.

23.34　$(24+6x)/(24-18x+6x^2-x^3)$.

第二十四章

24.40　$a_0=1.6, a_1=-0.8472, a_2=0.0472, b_1=0.6155, b_2=-0.1454$.

24.42　$a_0=2, a_1=-1, a_2=a_3=0, b_1=\sqrt{3}/3, b_2=0$.

24.43　$0.8; 0.8-0.8472\cos(2\pi x/5)+0.6155\sin(2\pi x/5)$.

24.45　$T_0(x)=1; T_1(x)=1-\cos(\pi x/3)+(\sqrt{3}/3)\sin(\pi x/3)=y(x)$.

24.46　$[(\sqrt{2}+2)/2]\sin(\pi x/4)+[(\sqrt{2}-2)/2]\sin(3\pi x/4)$.

24.47　$1-\cos(\pi x/2)$.

24.49　$\pi^2/12$ 与 $\pi^2/6$.

24.50　$\pi^2/8$.

24.52　$\pi^3/32$.

24.56　$1-\omega^2, 0, 1-\omega$.

24.57　$V^T=(3,-2,0,-1,0,-2)$.

24.58　$V^T=(5,1,5,1,-3,1,-3,1)$.

第二十五章

25.51　约为 1.839.

25.52　2;3;0.567143.

25.53　1.83929.

25.54　1.732051.

25.55　1.245731.

25.60　1.618034.

25.69　$x=0.772, y=0.420$.

25.72　3 与 −2.

25.74　$x^2+1.9413x+1.9538$.

25.75　4.3275.

25.76　1.123106 与 1.121320.

25.77　1.79632.

25.78　0.44880.

25.79　1.895494267.

25.80　$-0.9706 \pm 1.0058i$.

25.81　$x=7.4977, y=2.7687$.

25.82　$x=1.8836, y=2.7159$.

25.83　0.94775.

25.84　$x=2.55245$.

25.85　1.4458.

25.86　$x=1.086, y=1.944$.

25.87　1.85558452522.

25.88　0.58853274.

25.89　$(x^2+2.90295x-4.91774)(x^2+2.09705x+1.83011)$.

25.90 1.497300.

25.91 7.87298, −1.5, 0.12702.

25.92 1.403602.

25.93 1.7684 与 2.2410.

第二十六章

26.86 精确解为 0.8, 0.6, 0.4, 0.2.

26.88 精确解在题 26.55 中给出.

26.91 精确解为 5, −10, 10, −5, 1.

26.92 精确逆为 $\begin{bmatrix} 5 & 10 & 10 & -5 & 1 \\ -10 & 30 & -35 & 19 & -4 \\ 10 & 35 & 46 & -27 & 6 \\ -5 & 19 & -27 & 17 & -4 \\ 1 & -4 & 6 & -4 & 1 \end{bmatrix}$.

26.96 精确逆为 $\begin{bmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{bmatrix}$.

26.101 $2160\lambda^3 - 3312\lambda^2 + 381\lambda - 1 = 0$.

26.109 $(0, -i, i)$.

26.110 $\begin{bmatrix} 0 & i & 1 \\ -i & -1 & i \\ 1 & -i & 0 \end{bmatrix}$.

26.119 2.18518, −0.56031, 2.00532, −0.36819.

26.120 1.62772, 3, 7.37228.

26.121 8.3874, C(0.8077, 0.7720, 1); 4.4867, C(0.2170, 1, −0.9473); 2.1260, C(1, −0.5673, −0.3698); C 为任意常数.

26.122 $\begin{bmatrix} 5 & -10 & 10 & -5 & 1 \\ -10 & 30 & -35 & 19 & -4 \\ 10 & -35 & 46 & -27 & 6 \\ -5 & 19 & -27 & 17 & -4 \\ 1 & -4 & 6 & -4 & 1 \end{bmatrix}$.

26.123 $\frac{15}{64}\begin{bmatrix} 15 & -70 & 63 \\ -70 & 588 & -630 \\ 63 & -630 & 735 \end{bmatrix}$.

26.124 $\frac{1}{6}\begin{bmatrix} 6-8i & -2+4i \\ -3+10i & 1-5i \end{bmatrix}$.

26.125 98.522.

26.126 12.054; $[1, 0.5522i, 0.0995(3+2i)]$.

26.127 19.29, −7.08.

26.129 0.625, 1.261, 1.977, 4.136.

26.130 0.227 = 最小的.

26.131 否.

第二十七章

27.18 $(0,0), (0,1), \left(\frac{2}{3}, \frac{5}{3}\right), (2,1), (3,0)$; 极小值在 $\left(\frac{2}{3}, \frac{5}{3}\right)$ 处; 极大值在 $(3,0)$ 处.

27.19 参看题 27.18.

27.20 $-4y_1 - y_2 - 3y_3 =$ 极大; y_1, y_2, y_3 非负; $-y_1 + y_2 - y_3 \leqslant 1$, $-2y_1 - y_2 - y_3 \leqslant -2$.

27.21 参看题 27.18.

27.22 $4y_1 + y_2 + 3y_3 = \min.$; y_1, y_2, y_3 非负; $y_1 - y_2 + y_3 \geqslant 1$, $2y_1 + y_2 + y_3 \geqslant -2$; 解位于 $(0,0,1)$.

27.23　参看题27.18与27.20.

27.24　$x_1=\frac{3}{5},x_2=\frac{6}{5}$.

27.25　极端解点为(0,1)与$\left(\frac{2}{3},\frac{5}{3}\right)$.

27.27　支付为2.5;$R\left(\frac{1}{2},\frac{1}{2}\right),C\left(\frac{1}{4},\frac{3}{4}\right)$.

27.30　$\frac{37}{16}+\frac{17}{12}x+\frac{15}{8}x^2+\frac{1}{12}x^3$;1.3125;　2,　1,0,1,2.

27.31　$1.04508-2.47210x+1.52784x^2$;0.04508;0,.08,0.31,0.73,1.

27.32　同样结果;极大误差的5个位置.

27.33　极大=4.4　当 $x=(4.4,0,0,0.6)$.

27.34　极小$(5y_1+2y_2)=4.4$.

27.35

A	0	3	6	9	12
Max.	0	2	2	10	10

27.36　$\frac{3}{8},\frac{5}{8}$.

27.37　$R\left(\frac{1}{3},\frac{2}{3}\right),C\left(\frac{2}{3},\frac{1}{3}\right)$.

第二十八章

28.11　$x_1=3.90,x_2=525$, error$=0.814$.

28.12　$\rho=0.814,|\rho|_{max}=1.15$.

28.16　$x_1=-0.3278=x_2$,误差$=0.3004$.

28.17　$x_1=-\frac{1}{3}=x_2$.

28.18　3.472,2.010;1.582;426.

28.19　平均值为$(\sum a_i)/N$.

28.20　$x=(A+C-D)/3,y=(B-C+D)/3$.

28.21　$x_i=A_i+\frac{1}{3}(\pi-A_1-A_2-A_3)$.

28.22　$L_1^2=A^2-D,L_2^2=B^2-D,H^2=C^2+D$ 其中 $D=\frac{1}{3}(A^2+B^2-C^2)$.

第二十九章

29.46　$c_0=\frac{1}{15},c_1=\frac{1}{6}$.

29.52　0.2,0.5.

29.53　$T(x,y,z)=\frac{1}{6}[T(x+h,y,z)+T(x-h,y,z)+T(x,y+h,z)+\text{etc.}]$.

29.54　$y=(x-1)^n$.

29.55　一个近-奇异点在 $x=0$ 处.

29.56　$y=(1-e^{-nx})/(1-e^{-n})$.

29.57　一个近-奇异点在 $x=0$ 处.

29.58　精确解为 $1-\sqrt{2/\pi}\int_0^{x/\sqrt{\varepsilon}}[\cos(u^2)+\sin(u^2)]du$.

第三十章.

30.10　理论值为2,3,4与5个步长.

30.11　精确值为2.

30.14　理论值为$\frac{1}{16},\frac{9}{16},\frac{6}{16}$.

30.15　理论值为0.2401,0.4116,0.2646,0.0756,0.0081.

30.17　当 $N\to\infty$理论值为$1/e$.

30.18 理论值为$\frac{1}{4}$,$\frac{1}{2}$,$\frac{3}{4}$.

30.19 $y=-\log(1-x)$或同样地好 $y=-\log x$.

30.20 理论值为 0.3413,0.1359,0.0215,0.0013.